USER PERCEIVED VIDEO QUALITY MODELLING ON MOBILE DEVICES FOR VP9 AND H265 ENCODERS

Yao Xiao
Bachelor Degree of Information Technology

Submitted in fulfilment of the requirements for the degree of
Master of Information Technology (Research)

Science and Engineering Faculty
Queensland University of Technology
2015

Keywords

VP9, H264, H265, video encoder, codec, Subjective Video Quality Assessment (sVQA), Objective Video Quality Assessment (oVQA), Quality of Experience (QoE), user experience, video encoder performance, small factor screen, prediction model, predictor

Abstract

The function and hardware capability of mobile devices have been evolving at a staggering pace over the last few years, making it possible for end users to consume video contents on mobile devices with small-form factor screens. Unlike watching video contents on conventional devices such as TV, this change in user behaviour is creating new challenges for video content distributors and Internet service providers (ISP) because video contents take up large amounts of bandwidth,and therefore is exerting pressure onto the existing infrastructures. In order to alleviate the adverse situation caused by this rising demand for viewing videos online with mobile devices, a new generation of video encoders such as VP9 and H.265/AVC are being developed with the aim of reducing the bandwidth required while maintaining the same level of user-perceived quality. From the perspective of both ISP and video content distributors such as YouTube, encoding the video contents using the minimum bitrate with the most desirable encoders, while ensuring the quality of experience (QoE) of end users, will significantly reduce their operation cost.

The performance of VP9 and H.265/AVC encoders was analysed by both subjective and objective methods against the previous generation encoder, H.264/AVC. Test video sequences encoded by all three video encoders at different levels of distortions (bitrates) were evaluated. The outcome of both methods showed consistency and revealed that the latest generation of video encoders are about twice as efficient as H.264/AVC.

In this study, I have created prediction models by conducting subjective (user study involves participants) and objective analysis to predict the user perceived video quality on small-form factor screen. The prediction models take the definitions of video encoders, contents, and encoding parameters as predictors to estimate the user-perceived video quality. The prediction accuracy is determined on how accurately the average subjective scores gathered in this study can be estimated by the predictors. I have achieved prediction accuracy of 91.5% for the model that does not take objective scores as predictors and 94.5% for the model that does. Each proposed model takes 4 predictors.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACR	Absolute Category Rating
ACR-HR	Absolute Category Rating with Hidden Reference
ANOVA	Analysis of Variance
ATSC	Advanced Television Systems Committee
AVC	Advance Video Coding
BR	Bitrate
CB	Coding Block
CBR	Constant Bitrate
CD	Content Definition (Video)
CF	Coding Format
CI	Confidence Interval
CRT	Cathode Ray Tube
CTU	Coding Tree Unit
DMOS (ACR-HR)	Differential Mean Opinion Score
DMOS (DCR)	Degradation Mean Opinion Score
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
DVB	Digital Video Broadcasting
EBU	European Broadcasting Union
FR	Frame rate
FR	Full Reference
GOP	Group of Pictures
JCT-VC	Joint Collaborative Team on Video Coding
JM	Joint Model

JTC	Joint Technical Committee
JVT	Joint Video Team
HD	High Definition
HEVC	High Efficiency Video Coding
HVS	Human Visual System
IEC	International Electrotechnical Commission
Inter-Frame	P-Frame / B-Frame
Intra-Frame	I-Frame
ISO	International Organization for Standardization
ISP	Internet Service Provider
IRCCyN	Institut de Recherche en Cybernétique de Nantes
ITU	International Telecommunication Union
ITU-R	International Telecommunication Union Radio Communication Sector
ITU-T	International Telecommunication Union Telecommunication Standardization Sector
PCC	Pearson Correlation Coefficient
PSNR	Peak Signal-Noise-Ratio
QoE	Quality of Experience
MB	Micro-block
MMSPG	Multimedia Signal Processing Group
MOS	Mean Opinion Score
MPEG	Moving Picture Expert Group
NR	No Reference
NTIA	The National Telecommunications and Information Administration

oVQA	Objective Video Quality Assessment
QP	Quantization Parameter
RR	Reduced Reference
RS	Resolution
SB	Super-block
SI	Spatial Information
SR	Spatial Resolution
SROCC	Spearman Rank Order Correlation Coefficient
SSCQE	Stimulus Continuous Quality Evaluation
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation
SSIM	Structural SIMilarity
sVQA	Subjective Video Quality Assessment
TI	Temporal Information
TS	Transport Stream
UHD	Ultra High Definition
VBR	Variable Bitrate
VCEG	Video Coding Experts Group
VHS	Video Home System
VQA	Video Quality Assessment
VQEG	Video Quality Expert Group
VQMT	Video Quality Measurement Tool
4K	3840 × 2160 pixels video resolution
8K	7680 × 4320 pixels video resolution

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

QUT Verified Signature

Signature:

Date: 12 Feb 2015

Acknowledgements

This study would not have been possible, let alone completed without the assistance from various individuals and organizations. First of all, I would like to acknowledge the Singapore-based company, Amalgamated Leisure Pte Ltd, which has provided me with financial support throughout this Master by Research course of study at Queensland University of Technology. The general manager, Mr. John Tay and the stakeholder of the company, Dr. Winner Lim, have also provided me with important guidance as mentors. Their guidance and advice allow me to recognize both the intangible benefits of higher education and the importance of scientific research and academic achievement to the wellbeing of an individual in society.I would also like to acknowledge my principal supervisor, Dr. Wei Song, who has provided me with not only her consistent support and guidance of her superb expertise and in depth knowledge of my research field, but also the encouragement and spiritual support which was a great deal of help during my research. It is also important to mention that she had occasionally sacrificed her own personal time on weekends to do the editing of my proposed publications and verify the reliability of collected experiment data. Her level of sense of responsibility as my principal supervisor is phenomenal and therefore, I hereby sincerely express my utmost gratitude and respect for the level of work ethic she has displayed.

Special thanks also go to my associate supervisor, Assoc. Prof. Dian Tjondronegoro, who has overseen my research project from a higher level. He provided me with important advice in the big picture rather than the technical details, which helped me to effectively manage my time and resources in order to complete the research goals. Without his input, this study would not have been completed within the designated timeframe set by the University and extra resources might have had to be allocated.

Last but not least, I would also like to thank my friends and colleagues from the QUT Mobile Innovation Lab. Prithwi, Tony, Jimmy, and many more have shared joy, friendship and understanding with me throughout my course of study. All these friends and supervisors have made my course of study in QUT a quite unique and enjoyable part of my life journey.

Chapter 1: Introduction

This chapter outlines the motivation and background (section 1.1), the research questions to be addressed (section 1.2), and the research purposes and scopes (section 1.3). Section 1.4 describes how this study will supplement the knowledge of video encoder performance, and contribute to the field of study and justify its importance. Section 1.5 summarizes the remaining chapters of this document.

1.1 MOTIVATION AND BACKGROUND

Through the technological advancement in the past few years, the screen resolution of mobile devices has increased dramatically. Many mobile devices in the market come with High Definition (HD) screens of 1280 × 720 and full HD resolution of 1920 × 1080 pixels or better. Apple iPad Air, for instance, has a retina screen of 2048 × 1536 pixels (Apple Inc., 2014a). The latest mobile device, iPhone 6 Plus, comes with a native full HD screen (Apple Inc., 2014b). This advancement in mobile hardware display technologies has triggered user need of consuming high-resolution video content on mobile devices, and expanded their expectation of the quality of the video contents. Meanwhile, video content providers and distributors are actively adopting HD as a resolution standard; therefore Standard Definition (SD) video contents with spatial resolution of less than 720 × 576 pixels are rapid diminishing. Since HD resolution has become the mainstream, standards beyond it, such as UHD (Ultra High Definition) 3840 × 2160 (4k) and 7680 × 4320 (8K) resolutions, have already rolled out and 4K-capable display devices are already available in the market (Gong et al., 2012; Kim et al., 2009). According to a report (Cisco, 2013), mobile video traffic exceeded 50% of the total mobile data consumed in 2013. It was also estimated that the majority of video contents would be consumed on mobile devices such as tablets and smart phones on data network instead of on linear television (Ooyala, 2013). This new demand and industry trend consequently have placed the network infrastructures of Internet Service Providers (ISPs) under strain.

As a means of alleviating the increasing load of network infrastructure and cater for the rapidly expanding video resolution, sophisticated state-of-the-art video content compression methods such as AVS2, VP9 and H.265/HEVC (High Efficiency Video Coding) are being developed by well-established institutions such as ISO/IEC Moving Picture Experts Group

(MPEG), ITU-T Video Coding Experts Group (VCEG) and Google Inc. The latest generation of video encoders is designed to reduce the required bitrate of high-resolution video contents while ensuring end user quality of experience (QoE). Therefore, they are the potential antidote of the problems and challenges we are facing now.

Previous studies were carried out to analyze the performance of H.264/AVC encoder on conventional viewing device. However, due to the size of mobile device screen and the type of internet connection, researchers found the user experience (UX) could be very different from conventional linear TV systems (Knoche & McCarthy, 2004; Song, Tjondronegoro, & Docherty, 2012). New encoders are also expected to performance very differently as compared to H.264/AVC encoder. As revealed by subjective studies, both H.265 and VP9 encoders are shown to be capable of saving approximately half the bandwidth, compared to previous generation video encoders such as H.264/AVC and VP8 while maintaining same level of quality (Bankoski et al., 2013; Mukherjee et al., 2013; J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, & T. Wiegand, 2012). However, the actual performances of the latest encoders, as perceived by end users, are still unknown on small-form factor screens of mobile devices. The small-form factor screens are those 3 to 10-inch screens usually found on mobile devices. In particular, no study has been conducted to assess the performance of Google VP9 video encoder subjectively on small-form factor screen although there has been study carried out for the encoder on standard TV recently (Rerabek & Ebrahimi, 2014), to the best of the author's knowledge. This is perhaps due to the fact that VP9 is a relatively new codec and its adoption is therefore still limited at the time during this study was conducted. Moreover, a few objective studies of VP9 (Bankoski et al., 2013; Grois, Marpe, Mulayoff, Itzhaky, & Hadar; Mukherjee et al., 2013), gained inconsistent results. Very different from the over-50% of bitrate saving indicated in two studies (Bankoski et al., 2013; Mukherjee et al., 2013), a few studies (Grois et al.; Rerabek & Ebrahimi, 2014) suggested that VP9 was even inferior to H.264 with an average bitrate overhead of 8.4% at the same objective quality under certain situations. Hence, further research is needed to evaluate the performance of VP9 and to compare it with the latest H.265 encoder.

Furthermore, current research has focused mainly on oVQA methods to determine the performance of VP9 and H.265 encoders. The outcomes might be inaccurate and different from the user-perceived quality. The subjective video quality assessment (sVQA) method, commonly known as the most accurate method to gauge video quality, is sometimes used as a means to determine the effectiveness of objective video quality assessment (oVQA) methods

(Webster, Jones, Pinson, Voran, & Wolf, 1993). In other words, sVQA has to be carried out to supplement oVQA. Only by doing so, will researchers be able to gain insight about the human-perceived quality of the latest video encoders. If consistency exists between the evaluation results of subjective and objective studies of both H.264 and the latest encoders, researchers will be able to predict the subjective performance of the latest generation of video encoders by using their objective performance data or vice versa. This is significant, in that video content distributors will rely on the technologies that predict perceived video quality for network adaptive content provisioning once H.265/AVC and VP9 encoders are widely adopted by video content distributors.

Due to the rapid evolution of video content delivery infrastructure and digital hardware performance recently, the video industry needs to know about the true performance of the latest video encoder in close relation with User Experience (UX). Ensuring Quality of Experience (QoE) while delivering video contents with the minimum bandwidth will significantly save the operating cost of video content distributors and enable them to remain competitive in the industry. There is proven correlation between sVQA and oVQA for the previous generation of video encoders such as H.264/AVC (Vranjes, Rimac-Drlje, & Zagar, 2008). If such correlation could be proven existing in the latest generation of video encoders as well, video content distributor will only need to make minor modifications to their existing video content delivery network once H.265/AVC and VP9 are widely adopted. Since the adoption of the latest generation of video encoders is just a matter of time, any research addressing this knowledge gap will help video content distributors to devise their video delivery strategy online, improving the user experience of watching video online on small-form factor screens.

1.2 RESEARCH QUESTIONS

The main research question is as follows:

What is the user perceived performance of the latest generation of video encoders such as VP9 and H.265/AVC on small-form factor screens?

In order to solve the proposed research question, the following subordinate research questions are drawn to break this research project into manageable chunks.

- What assessment methods can be used to measure the performance of VP9 and H.265/AVC encoders subjectively (sVQA) and objectively (oVQA) on mobile devices?

- How much performance improvement can the latest generation of video encoders achieve on small-form factor screens, compared to H.264 encoder based on sVQA and oVQA?

- What are the influential factors affecting the VQA results and the correlations between them?

- How can we predict the sVQA and oVQA outcomes for any given video content compressed by the latest generation of video encoders, based on the influencing factors?

1.3 PURPOSES AND SCOPE

A vast number of factors will affect the user-perceived quality of video contents. Such factors include the distribution network condition, the distortion caused by compression, the viewing condition and the psychology of participants. Regardless which approach of video quality assessment, research scope must be set to align to real life situation where there will be constraints. Furthermore, clearly defined scope will help to focus the study.

First of all, the definition of the video content discussed and tested in this study refers to 2D videos broadcasted by video content distributors, which typically include sports, news, movie, drama, variety show and animation. They are natively produced and broadcasted in HD or full HD without any up-scaling and up-conversions.

Secondly, the video content distortion described in this study only refers to the distortion caused or introduced by video encoders, without touching on other factors which are irrelevant to encoding technologies, such as distortions caused by mobile network signal, network bandwidth and psychological factors of the participants.

Thirdly, this study aims to discover the realistic performance of the latest generation of video encoders of VP9 and H.265 with both subjective and objective VQA methods. A previous generation of video encoder, H.264, was tested for comparison purposes. Other video encoders were not tested in this study.

Lastly, due to technical constrains, H.265 video sequences could not be played back on the tablet device that was used in this study. Therefore, the sVQA prediction models were created based on the sVQA scores of VP9 and H.264/AVC encoders and on the oVQA scores of all three encoders.

1.4 SIGNIFICANCE AND CONTRIBUTION

The prediction models created in this study will provide an automated and cost effective means for video content distributors to deliver their service online while ensuring positive user experience on small-form factor screens. In a real life situation, sVQA cannot be conducted due to many constraints such as cost and time. Video content distributors have to rely on automated objective video quality assessment methods to achieve adaptive bitrate video streaming in real time. However, objective assessment scores are not closely associated with the user perceived quality. On the contrary, the prediction models can be embedded into the automated online video content evaluation mechanisms currently used by the video content distributors, allowing such mechanisms to predict the user perceived video quality cost effectively instead of via objective scores. Video encoding parameters can be adjusted automatically to achieve a balance between user experience, encoding bitrate, encoding time and so on and such factors. The higher the percentage of chance the model can successfully estimate the subjective scores garnered in this study, the better the model. Objective scores such as SSIM and PSNR able to be used as predictors to further enhance the prediction accuracy.

Besides their improved relevance to perceived quality, the prediction models also specifically focus on the user experience on small-form factor screens. As described in Chapter 1, Section 1.1, the increasing change of user behaviour away from watching video on conventional TV is shaping the user experience in a new way. The perceived video quality on conventional TV is speculated to be different from that on small factor screens. The prediction models provide specific ways to address this issue.

From the end user perspective, the realistic performance of the latest generation of video encoders estimated by the prediction models will give an accurate preview of how much quality improvement new video encoding technologies can bring to the existing ones, allowing them to voice their opinions about it. These opinions will influence how the latest generation of video encoders will be fine-tuned and developed to suit the demand of end users in a macro point of view. As such, user demand drives technology while technology evolves for the better according to end user expectation.

From the perspective of the academic world, effective estimation of the realistic performance and QoE of the latest video encoders via performance prediction models will help researchers to improve on the conceptual or theoretical designs of modern video encoding technology. Since the original design of the first generation video encoders,

researchers and scientists have been making incremental designs to improve the performance of video encoders in the past decades based on user feedback and market demands. The prediction models are also expected to allow both scientists and researchers to simulate the final expected performance of any video encoders that are under development. This is essential in the sense that our business environment and user demand are evolving in a staggering pace nowadays due to the rapid advancement of information technology and globalization. Such a phenomenon is driving new technologies to be developed and implemented in short cycles. The performance prediction models of the latest encoder provides an important interface for technologies beyond H.265 and VP9.

1.5 THESIS OUTLINE

Chapter 2 aims to organize the literature related to the proposed research. Chapter 3 illustrates the technical details of how the subjective and objective studies are conducted. Chapter 4 lists out the study outcome based on the data collected. In Chapter 5, data analysis will be carried out on the subjective and objective scores collected. Performance prediction models are proposed to estimate the user-perceived quality of videos encoded by the latest generation of encoders. The last Chapter summarizes the research outcome and discusses the implications of this study and how the study outcome fulfills the designed research goals.

Chapter 2: Literature Review

This chapter aims to define, critically analyse and categorize the background information and knowledge related to the identified research problem by examining the existing literature. Knowledge, ideas and philosophies from these literature, that are helpful to design, structure and carry out the research, will be discussed. In the process of accessing these materials, clues of what have already or have not been done by other researchers will be sought to further discover the research gap.

The interconnections in the existing related literature will be established in this document through cross referencing. Due to the fact that this study will cover vastly different knowledge, ranging from statistical data analysis to video compression technologies, all knowledge will be categorized and their connections will be established in this chapter.

The literature review will start with a background overview of the contemporary moving picture industry to discuss about how market and consumers have driven it to develop new generations of video encoders in the past decades and the prospective of the industry. The latest video encoders from the established organizations such as ITU and Google Inc. will then be discussed and analyzed technically based on existing literature. The approaches, methods and philosophies to evaluate these encoders will be covered. Lastly, data analysis and modeling techniques for video codec performance will be discussed.

2.1 DISPLAY DEVICES AND RESOLUTION

Based on the forecast revealed in the Cisco Visual Networking Index (VNI), the expected global mobile data traffic will reach a staggering 11.2 exabytes per month, mainly resulted from video material consumption by 2017, an almost 13-fold increase since 2012. The total number of global 4G connections will increase more than 16 times, reaching a figure slightly less than 1 billion (Cisco, 2013). Compared to conventional TV, video quality on mobile devices is confined by the hardware capabilities of the viewing devices and the network condition. Earlier researches (Jumisko-Pyykkö & Häkkinen, 2005; Knoche, McCarthy, & Sasse, 2005; Ries, Nemethova, & Rupp, 2007) are mostly out-dated and cannot be used as accurate references since the mobile device industry has taken off and expanded tremendously over the last few years. With the rapid hardware advancement of mobile devices, their native screen resolutions now reach full HD or better (SGP-NewsMan, 2013). On the other hand, the

mobile devices used in previous researches often came with native screen resolution far lower than full HD. Additionally, mobile phone networks have been undergoing significant technological transformations as video content distributors such as YouTube are already supporting full HD video playback on mobile devices.

Besides the lower resolution of small-form factor screens used in prior studies, the video sequences tested in these studies have resolutions far lower than HD as well. However, producing and recording video contents in full HD or higher is gradually becoming the standard practice in video production industry (Bankoski et al., 2013; Mukherjee et al., 2013). Hence, these studies are very limited, as they do not provide a context relevant to the prevalent higher video resolutions. Additionally, video content distributors are actively distributing content in high definition and there are academics suggested the resolution of video contents should match the native resolution of display hardware to achieve the most desirable QoE (Cermak, Pinson, & Wolf, 2011). The video contents of lower resolution than HD tested in previous studies are therefore not suitable to be tested on the latest mobile devices, which come with screens of much higher native resolutions. Assuming the popular resolutions of 720p and 1080p are adopted in both subjective and objective VOA methods, the mobile device chosen for sVQA should have native screen resolutions of HD or full HD.

Existing study also suggests the human vision system (HVS) is sensitive to contrast and sharpness (Ibrahim Ali, 2007) and it is logical to assume video sharpness can be easily enhanced by increasing the resolution (total number of pixels). Therefore improved screen resolution of both mobile device screens and video contents are expected to change user perceived video quality dramatically. There is no existing sVQA study that focuses on HD and full HD videos on mobile devices with native screen resolutions of full HD or better.

2.2 VIDEO COMPRESSION

In the 1980s and 1990s, video contents were stored on analogue storage mediums such as magnetic tapes called Video Home System (VHS) (Boucher, 2008). Due to cost, size and portability issues, Sony, Phillips, Matsushita and JVC standardized the Video Compact Disc (VCD) in 1993 (Schylander, 1998) for storing video contents digitally. Although digital video signals stored on disc are superior to the analogue signals stored on VHS, only a few minutes of raw videos could be fitted into one piece of disc. Since the signal of digital video contents is binary based, engineers then developed mathematical video compression methods to reduce

the signal size. As a result, a VCD is capable of storing 80 to 90 minutes of video content by adopting the MPEG-1 video compression arithmetic.

The rapid development of information technology has enabled the video industry to go much further. The rising complexity and performance of encoding and decoding hardware have stimulated the need for higher resolution videos. As a result, video compression methods have evolved through four generations in over the last 20 years (see Table 1). H.265/AVC and VP9 are the latest (4th) generation of video encoders.

Table 1. Generations of video encoders

Generation	Encoder
1	MPEG-1, H.261/AVC
2	MPEG-2, H.263/AVC
3	MPEG-4, H.264/AVC, WMV, VP8
4	H.265/AVC, VP9

The performance improvement between the two generations of video encoders is typically 50% or better, according to previous studies (Pourazad, Doutre, Azimi, & Nasiopoulos, 2012; Smith, 2006; Wong & Chen, 1993). However, as the literature suggests, the tremendous performance gain between these two connecting generations is achieved not by radical change of fundamental theories and designs, but by making incremental improvements to existing encoder designs. For example, MPEG-1 already had the concept of macro-block (MB), which has been improved for H.265/AVC and VP9 encoders.

MPEG1 and H.261 can be seen as the first generation of video encoders, developed in the early 1990s; MPEG2 and H.263/AVC are the second generation, developed short after the first generation; MPEG4, H.264 and VP8 are the currently adopted third generation video encoders; VP9 and H.265/AVC are the latest generation video encoders, which were finalized in 2013 and still under development.

2.2.1 State-of-the-Art Video Encoders

In video compression, the video encoder defines how the original video signal, represented in digital binary form, is compressed and stored as arrays of pixels by using mathematical algorithms. Although differencing in minor details, the fundamental design of modern encoders follows the main principle of the block-based hybrid encoding approach by breaking down a frame of image into smaller blocks (Sikora, 1997; Sullivan, Ohm, Han, & Wiegand, 2012; Wiegand, Sullivan, Bjontegaard, & Luthra, 2003).

Uncompressed digital video signals consist of periodic sequences of images, referred to as frames, that are described by a few parameters including spatial resolution, colour space, colour sampling and bit-depth. Besides these attributes, compressed video content will have the bitrate as an additional frequently referred attribute, which represents the amount the binary information is used to construct the video frames that are displayed in one second. Each frame has a two-dimensional array of pixels that contains brightness and colour information (YUV)(Chen, Kao, & Lin, 2006). Uncompressed video signals usually contain rich brightness and colour information that need to be partially filtered out before or during the actual compression process. It is however possible that the original video signal is already sub-sampled by cameras or recording devices.

Before any motion information and video frame are compressed, they are put into the luma-chrominance colour space. Modern encoders separate video signals into three components for colour representation: Y, Cb and Cr respectively. The Y component, *luma or luminance,* represents brightness; Cb and Cr, *chroma* components, depict how much a colour is different from grey to blue and red. Typically, each component is represented by 8-bits precision. Interpolation of all three YCbCr components discards some of the *chroma* information to take advantage of the nature of the HVS (Dumic, Mustra, Grgic, & Gvozden, 2009). More bits can be saved without causing significant image quality deterioration to human observers, as the HVS is far less sensitive to colour than to brightness.

H.264/AVC Encoder

The H.264/AVC encoder was standardized by both VCEG and MPEG in 2003 (Richardson, 2004; Wiegand et al., 2003). The H.264/AVC encoder was designed to reduce the bitrate by half while maintaining the same level of image quality, compared to H.263 and MPEG2 standards (Kamaci & Altunbasak, 2003). Since its finalization, H.264/AVC encoder has gained popularity in the video industries, becoming the standard of delivering HD video contents over the Internet during the past decade. This is partly due to its close association with previous standards such as H.263, because many of its features are derived from H.263. Previous study indicates the reason that the H.264/AVC encoder quickly gained favor is partly because the video industries conservative and prefer to take one step at a time rather than changing to a completely new design (Richardson, 2004).

After the standardization, various parties developed their own versions of H.264 encoders short after, whether they were loyalty free or not. The open source implementation of the standard is x264 (Merritt & Vanam, 2006). The well-known open-source video encoding /

decoding tool, FFmpeg (Bellard & Niedermayer, 2012) ,with libx264 library, is widely used for both research and commercial purposes. FFmpeg is widely used by many applications such as ffdshow and MEncoder and is proved to be reliable and robust. Previously studies showed the x264 encoder outperformed not only a few commercial encoders (Vatolin, Kulikov, Parshin, Titarenko, & Soldatov, 2007) but also the Joint Model (JM) encoder (Model, 2008) developed by Joint Video Team (JVT) (Merritt & Vanam, 2006), which consisted video coding experts from ITU-T study Group 16 (VCEG) and ISO/IEC JTC 1 SC 29 / WG 11 (MPEG).

Google VP9 Encoder

The VP9 encoder is the successor to VP8, an encoder originally created by On2 to challenge with the dominant H.264/AVC encoders (Feller, Wuenschmann, Roll, & Rothermel, 2011). Google Inc. acquired On2 in 2009. Due to the unusual business strategy and model adopted by Google, the company made VP9 a freely available video encoder to everyone (Protalinski, 2013), officially releasing its VP9 video encoder on 12 June 2013 (http://www.webmproject.org/vp9/). Although the design of the VP9 encoder is already finalized, there was no official specification document available as pointed out by academics (Rerabek & Ebrahimi, 2014) (Řeřábek & Ebrahimi, 2014)at the time this study was conducted. The encoder and decoder contained in the package are complete.

Unlike other companies and organizations, Google Inc. has vast amounts of resources in hand and ready-to-go platforms such as YouTube to promote its technologies. This enables the VP9 encoder gained favor rapidly. Moreover, the resources Google has have already pushed the VP9 standard years ahead of H.265/AVC in term of completeness. H.265/AVC is still in the testing stage, taking its shape slowly, whereas working VP9 encoders, decoders and footages are already freely available online.

H.265 Test Model Encoder

The H.265/AVC encoder is designed to be the successor of the dominating H.264/AVC encoder. Just like the H.264/AVC, many designs and features of H.265 encoder are inherited from the H.263. For the H.256 encoder, we used the latest HEVC (High Efficiency Video Coding) Test Model HM14 by the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG). Compared to its preceding version, H.264/AVC, the H.265 encoder claims to be 50% more efficient. Previous studies revealed that H.265 encoder is extremely efficient for both random access and all intra configurations (Nguyen & Marpe, 2012). This shows the H.265 codec is not only highly efficient in video

encoding but also excellent for still image compression (Hanhart, Rerabek, Korshunov, & Ebrahimi, 2013). The first version of the test model encoder, finalized in 2013, was downloaded from the official repository at Fraunhofer Heinrich Herts Institute (HHI) via its original URL, https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/. The test model encoder comes with a few preconfigured profiles. Random Access (RA) profile outperforms other profiles such as Low Delay Profile (Hanhart, Rerabek, De Simone, & Ebrahimi, 2012). Since the Test Model HM14 encoder is still at its experimental stage, unlike the VP9 and H.264 encoders, it has a fixed-length hierarchical Group of Pictures (GOP) structure.

2.2.2 Technical Specifications of Video Encoders

Video sequences consist of a series of frames that are displayed at a typical speed of 25 to 30 frames per second. Analogue TV systems at different geographical locations displayed video frames in two different ways in the past. All the countries in the world were categorized under either PAL or NTSC regions mainly. Countries such as North America and Japan uses progressive scanning with 30 frame per second while commonwealth countries mainly stick to PAL interlaced scanning with 25 frames per second and two fields per frame (Haskell, 1997). Such a convention is still preserved in the third generation of video encoders, such as H.264/AVC and VP8 support field coding (interlaced scanning mode). However, scientists and researchers believed interlaced scanning is an inherited technologies of the obsolete technology and therefore ceased the field coding support in the latest generation of video encoders such as VP9 and H.265/AVC, because UHD contents are standardized, going to be produced, stored and transmitted by using progressive scanning mode exclusively (Sullivan et al., 2012). On the contrary, some academics believe the latest generation of video encoders will perform significantly better if field coding is supported (Henot, Ropert, Le Tanou, Kypreos, & Guionnet, 2013).

Regardless of the presence of field coding, modern video encoders such as H.264/AVC, VP9 and H.265/AVC divide any given frame into MB of different sizes for compression and processing. For the VP9 encoder, MB is commonly referred as super-block (SB) as each MB is commonly considered as a processing unit by video encoders. The components of a video sequence and the order of which they are processed are illustrated in Figure 1.

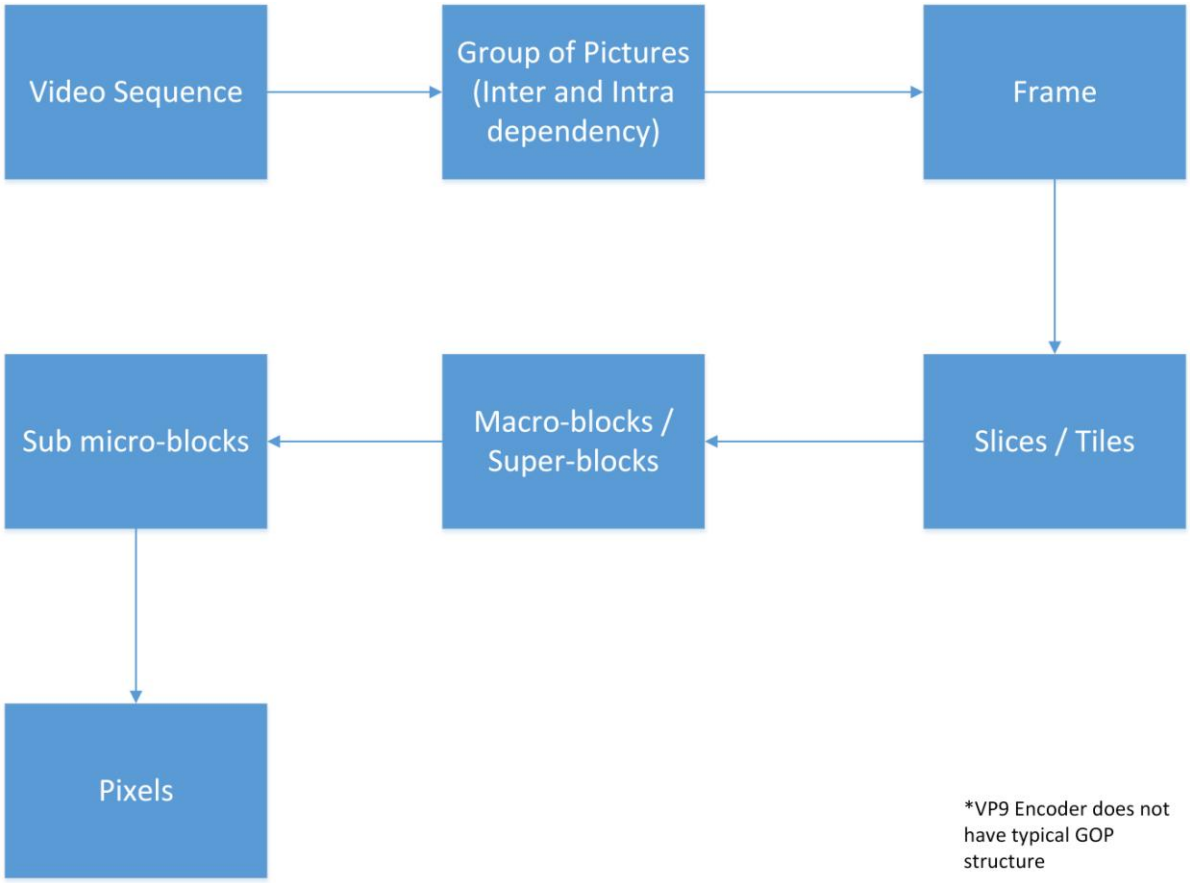


Figure 1. Components of video sequences

A chunk of digital video signal represented in binary is usually referred as a video sequence. When encoders process the video sequence, it is broken down into different units, level by level, for compression or processing. Considering above the binary level, the largest component of a video sequence is a GOP while the smallest one is a pixel. In an encoded video sequence, the GOP consists of a group of video frames that are interdependent representing motion when decoded by video decoders.

Every video frame consists of many MBs come with different sizes labeled by the number of pixels vertically and horizontally. Multiple MBs that remain unchanged or extremely similar in a series of video frames are commonly grouped by video encoders to form slices and tiles for bitrate saving proposes.

Figure 2. Intra Frame with Macro-blocks determined by VP9 encoders

Figure 3. Intra Frame with Macro-blocks determined by H.264/AVC encoder

Figures 2 and 3, generated by the CodecVisa bitstream analyzer, demonstrate the MB structure breakdown of a video frame for both H.264/AVC and VP9 encoders. As we can observe, the VP9 encoder has significantly fewer MB than the H.264.AVC encoder has. This is because the H.264/AVC encoder is unable to generate MBs bigger than 16x16 pixels and therefore it is observed that the total number of MBs required for H.264/AVC in a given video

frame is much larger than the total number of MB required for VP9 and H.265/AVC encoders, since they both support much larger MB sizes (64 × 64 maximum). For both video encoders, at regions that do not contain highly detailed image information, the blocks are briefly subdivided or not divided at all, whereas regions containing complicated texture, colour and brightness information are subdivided to more blocks. Such design improves the performance of encoders by allocating more bits to the information-rich regions and less to others (Chan, Yu, & Constantinides, 1990). Both 4th generation and H.264/AVC video encoders are designed this way. The major difference between them is the largest supported MB size.

There are also other minor differences between them. The previous generations of video encoders did not have the notion of a coding tree unit. H.265/AVC, for example, has the basic processing unit described as coding tree unit (CTU) which functions as super-block (SB) in VP9. Similar to SBs in VP9, CTU can be subdivided into coding blocks (CBs) in a so-called quad-tree structure (Choi & Jang, 2012; Sullivan et al., 2012). The scanning order for a given CTU or SB is from top left-hand corner to bottom right-hand corner as illustrated in Figure 4 and 5. Both VP9 and H.265/AVC allow 64 × 64 as the largest MB which can be further subdivided. For instance, a 64 × 64 SB can be subdivided into four 32 × 32 sub-MBs, which can be further split into four 16 × 16 SBs, identical to the largest SB size specified in the H.264/AVC encoder. However, unlike H.265/AVC, which has official documentation of technical specification, there are no official technical details about VP9 from Google Inc. Therefore, technical specifications of the VP9 encoder are extracted from existing publications and papers (Mukherjee et al., 2013).

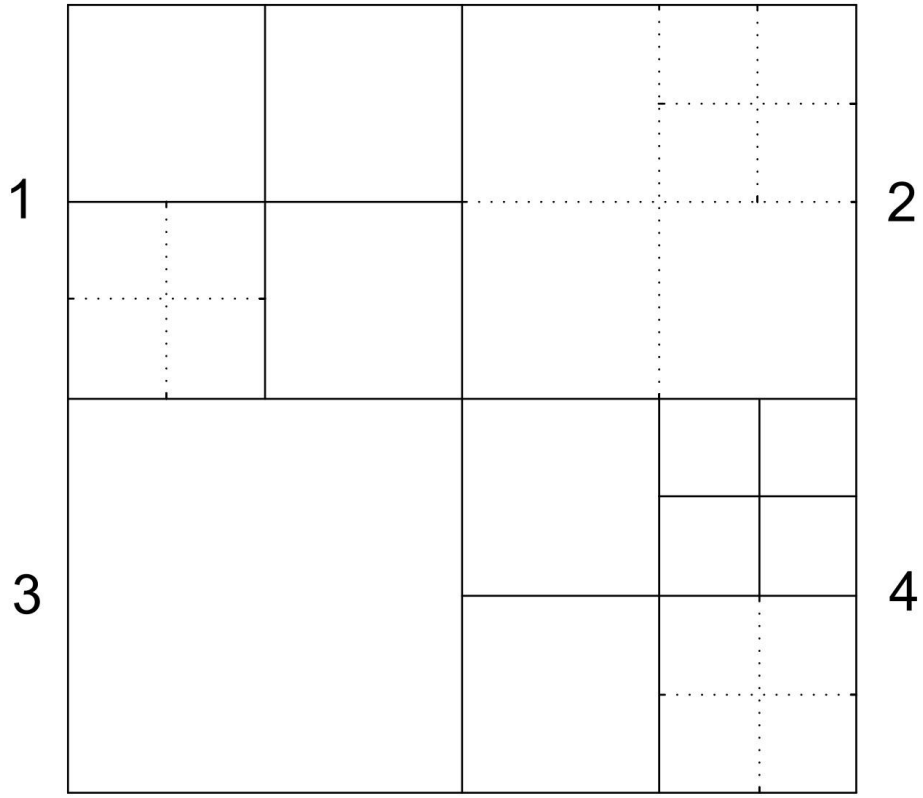


Figure 4. 64×64 pixels MB of H.265/AVC encoder

As shown in Figure 4, only square MBs are allowed by the H.265/AVC encoder. H.264/AVC separates a large MB into four quadrants of the same size and shape. When necessary, the H.265/AVC encoder can breakdown each quadrant to the smallest 8 × 8 MBs or 16 × 16 to interpret extreme image details and colour information.

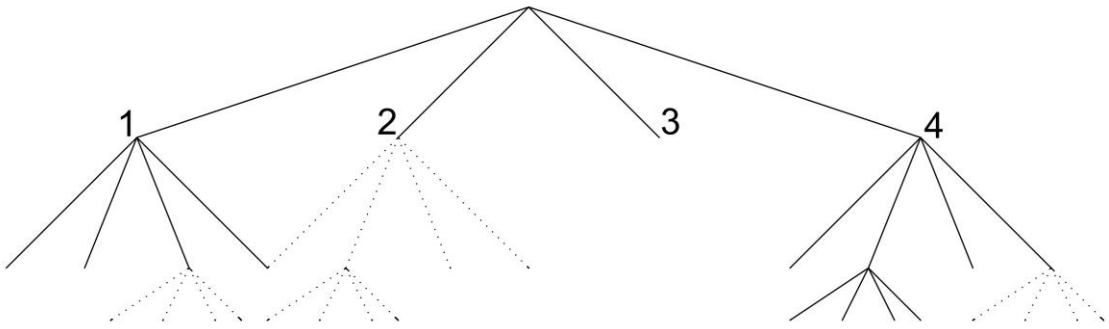


Figure 5. Quad-tree structure of H.265/AVC and VP9 encoders

Figures 4 and 5 show the MB structure of the H.265/AVC encoder and the order which each MB is processed. Sub MBs are processed from the top left hand corner to the bottom right-hand corner. A SB in VP9 or CTU in H.265/AVC can be subdivided into smaller MBs, and the processing order of for both encoders is identical.

Each MB contains brightness and color information (YUV) and each of the YUV components is processed individually. A combination of three factors is considered by modern video encoders to compress MBs – prediction, transformation and quantization, and entropy coding. Depending on the complexity of the image, modern encoders are capable of creating MBs of different sizes for the purpose of achieving better encoding efficiency. H.264/AVC, for example, is capable of creating 16 × 16 or 4 × 4 pixels MBs for *luma* and 8x8 MBs for *chroma* (Cb and Cr) information (Chen et al., 2006). However, the latest generation video encoders such as VP9 and H.265/AVC are capable of creating 64 × 64 MB (Sullivan et al., 2012).

Although such designs in 4th generation video encoders are counter-intuitive and seem to consider less about representing detailed video image, VP9 and H.265/AVC video encoders improve their performances significantly by saving more bits at static, and less detailed portions of a video frame, allocating more bits to image regions contain complicated details. Existing studies show forcing the H.265/AVC test model encoder to encode 16 × 16 MBs only, instead of 64 × 64 MBs will decrease the performance of the encoder by 11% (J. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, & T. Wiegand, 2012). One existing study also suggests having larger MBs in video sequences of higher spatial resolution will improve the performance of the video decoder significantly (up to 60%) by lowering the decoding complexity mathematically, fewer MBs will then be assessed by encoders (J. Ohm et al., 2012).

Modern video encoders predict and store the differences between one MB and another by using a motion vector. Only these differences, usually referred to as residual error data (excluding the actual MB), along with motion vectors are used to decode inter mode encoded MBs or frames. Two prediction modes of intra-frame (I-frame, spatial) and inter-frame (B or P-frame, temporal) prediction are available. Intra-frame prediction works within a given frame based on the encoded and decoded information of a particular MB, whereas, during inter-frame prediction, the encoder will take a reference MB to search for similar MB from nearby video frames in a sequence by using a block matching logarithm. Both motion vectors and residual error data will be generated during encoding. Figure 6 shows the inter-frame and intra-frame breakdown in a video sequence encoded by the VP9 encoder. Inter frame of a video contains both inter and intra MBs: Green MB represents inter MB; red MB is intra MB. Arrows represent motion vectors.

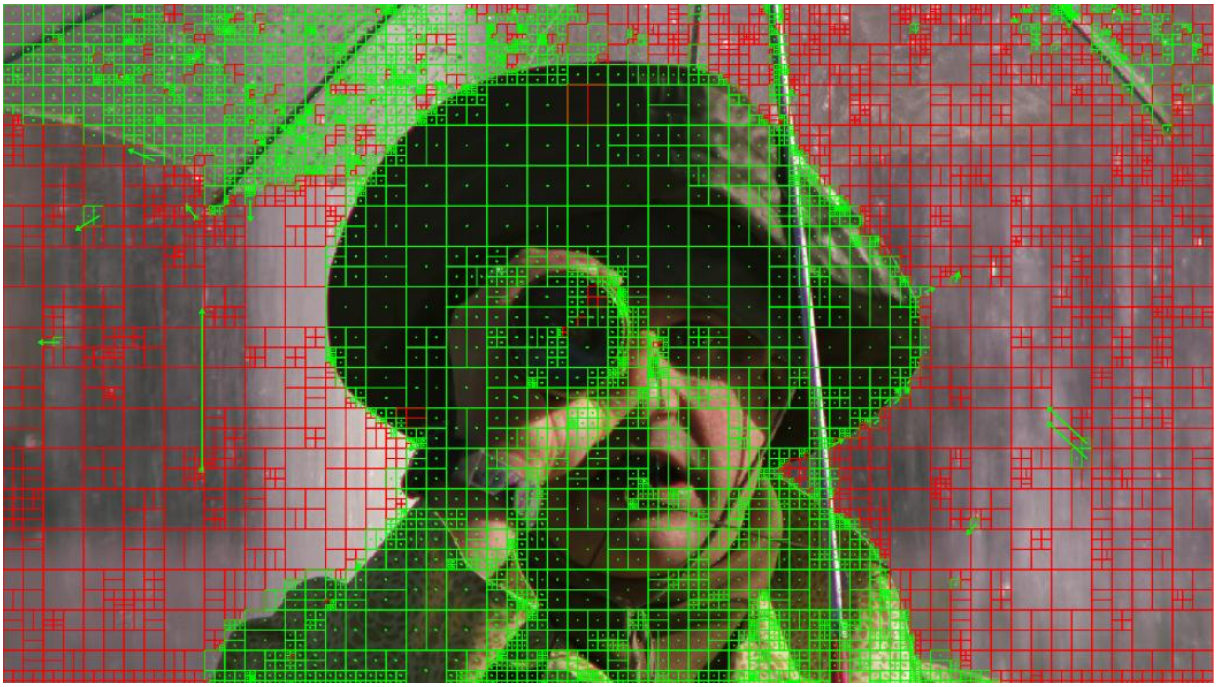


Figure 6. Inter Frame with MB determined by VP9 encoders.

All these three encoders follow the basic pattern of having MBs structure, motion vector and inter and intra frames or MBs in their designs. However, the 4th generations video encoders have minor design differences. For example, H.265/AVC has 33 directional intra predictions (Sullivan et al., 2012) that allow it to have a more accurate motion prediction of similar MBs, while the VP9 and H.264/AVC encoders support only 8 (Mukherjee et al., 2013). In contrast to the H.265/AVC and H.264/AVC encoders, the VP9 encoder supports rectangular MBs for potentially more versatile and effective breaking down of video frame into MBs. MBs and CTUs are defined as square for the H.265/AVC and H.264/AVC encoders only. Table 2 displays the difference in specifications of these three video encoders, as well as the sizes of MBs supported.

Table 2. Major differences in design of H.264/AVC, H.265/AVC and VP9 Encoders

	H.264	**H.265**	**VP9**
Supported Block Size	4 ×4, 4 ×8, 8 ×4, 8 ×8, 8 ×16, 16 ×8, 16 ×16	4 ×4, 8 ×8, 16 ×16, 32 ×32, 64 ×64	4 ×4, 4 ×8, 8 ×4, 8 ×8, 8 ×16, 16 ×8, 16 ×16, 16 ×32, 32 ×16, 32 ×32, 64 ×64
Directional Intra Prediction Mode	8	8+2 non-directional	33+2 non-directional
Support Slice	Yes	Yes	No
Support Tile	Yes	Yes	Yes
Presence of GOP structure	Yes	Yes	No
Presence of hidden reference frame	No	No	Yes

See Figure 7, the VP9 encoder is capable of creating MBs in the size of 4 × 8, 8 × 16 and 16 × 32. Such design is expected to improve the effectiveness of intra and inter prediction and reducing MB overhead, result in higher encoding compression efficiency as fewer subdivisions of MBs are necessary to represent a given image (Mukherjee et al., 2013). One of the unique features of the VP9 encoder, the hidden reference frame (super-frame), can be used as a reference frame just like inter-frame during playback. However, unlike an inter video frame, the hidden reference frame in VP9 is not displayed. Although such design might improve the encoding efficiency, it can also cause problems when fitting VP9 video streams into containers, since the purpose of video container is to display every video frame in a stream. Special designs of containers are required to cater for VP9 streams playback.

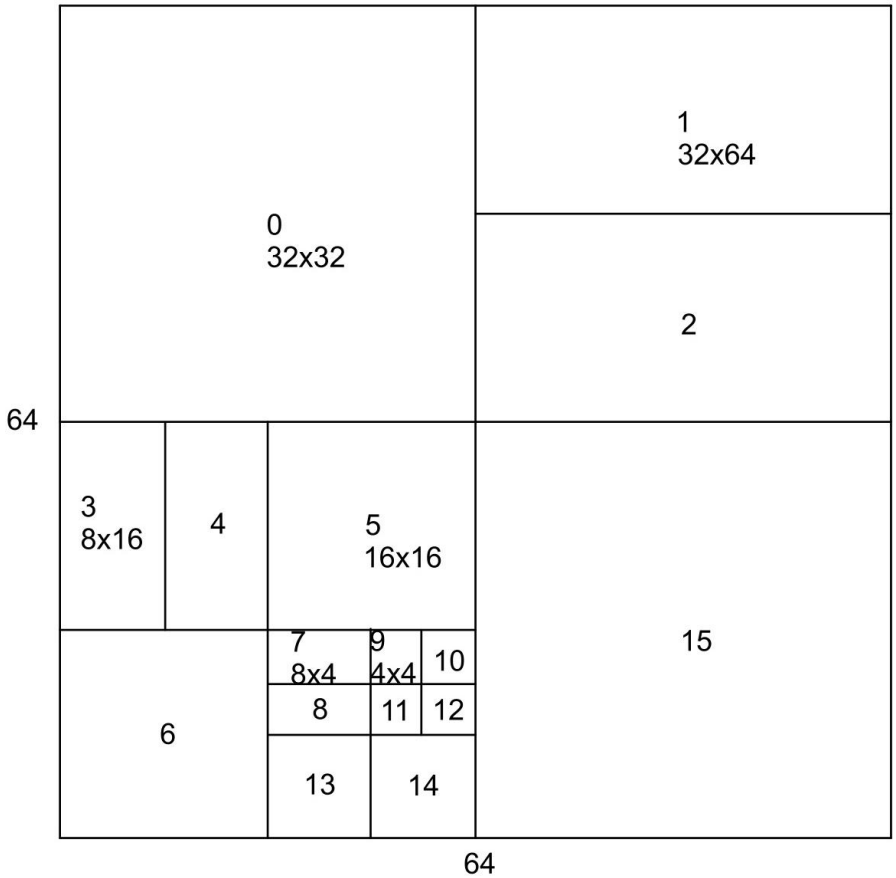


Figure 7. VP9 encoder SB sub-division.

Currently, VP9 streams can only be wrapped into IVF and WEBM (a simplified version of MKV) containers, whereas H.264/AVC streams have significantly more choices. Due to the fact that the H.265/AVC encoder is still a test model encoder, there was no container available for it at the time this study was conducted. However, MPEG is currently working on projects

to support the H.265/AVC video encoder in the MPEG transport stream (TS) used by ATSC (Advanced Television System Committee), DVB (Digital Video Broadcasting) and Blu-ray Disc and ISO base media file format such as MP4 (Standardisation, 2013).

Conventionally, video encoders produce a three types of frames: I, P and B frames. When they are grouped together to display a video sequence, they are commonly described to have a GOP structure (e.g. a typical GOP structure is IBBPBBPBBI). An I-frame can be viewed as an independent reference frame that allows skipping during viewing or playback and consequently uses a significantly larger amount of data than other types of video frames. The P-frame is more compressible than the I-frame and uses data from previous frames for decompression. The B-frame is the most compressible frame, as it is capable of using data from both previous and preceding frames for compression. Due to its heritage, the H.265/AVC video encoder has similar GOP coding structure as the H.264/AVC encoder, whereas the VP9 encoder is different from the aforementioned two and do not have the GOP structure. The VP9 encoder does not encode B-frames; instead, it supports hidden frames function similar to inter frames. The Hidden frame supplies additional decoding information for the decoding process (Mukherjee et al., 2013) and will not be displayed by a decoder. Having B-frames in a video sequence will cause frame reordering, therefore increasing the overhead of decoders, and placing strains on the encoding hardware. In other words, keeping the B-frames out will decrease the decoding complexity and lower the playback hardware requirement.

2.3 VIDEO QUALITY ASSESSMENT

End-user perceived video quality is often affected by external factors. To align with the research goals specified in Chapter 1, the existing literature on video quality assessment without considering external factors will be the primary focus.

A few main external factors have to be excluded from the literature review of video quality assessment. The quality of the video contents delivered through mobile network to portable devices is subject to different forms of distortions (Yuen & Wu, 1998). For example, the network condition will affect end-user perceived quality;also the hardware capability of the mobile devices used by end users varies and therefore will also significantly affect user experiences (Frojdh, Horn, Kampmann, Nohlgren, & Westerlund, 2006). Factors such as network condition and hardware capability or specification are excluded from this assessment.

This section focuses on four important aspects in video quality assessments: video content selection, encoder configuration, subjective and objective video quality assessment methods.

2.3.1 Video Content Selection

To accurately emulate the real-life situation of video content viewing by end users and to rule out the inconsistency that might arise during video quality assessment processes, various organizations have produced standardized video content databases for research purposes and made a number of specifically designed video databases (ITUR Rec, 2012b) that cater to various research purposes freely available to all researchers. A summary document titled QUALINET Multimedia Database produced by Czech Technical University in Prague is updated regularly to keep track of all the available databases (Fliegel, 2014). At the time our study was conducted, Version 5.0 of the document was referred.

Currently, there is a very limited amount of VQA research done on small-form factor full HD screen devices. Researchers are still in the stage of designing tools and new VQA methods that will be served as the cornerstones for future work. A large number of video databases exists, most designed with a specific purpose or function in mind (Winkler, 2012). For example, Poly@NYU Video Quality Database Packet Data Loss Database (Yang, 2008), as its name suggests, focuses on the quality evaluation of quality deterioration caused by data network and transmission factors. Similarly, LIVE Video Quality Database created by Laboratory for Image & Video Engineering of The University of Texas at Austin provides 10 different uncompressed high-quality reference video sequences for researchers to carry out VQA studies on MPEG2 and H.264/AVC encoders regarding network transmission loses, with the aim of accurately simulating real-life situations (Seshadrinathan, Soundararajan, Bovik, & Cormack, 2010).

Specifically devised video databases for carrying out research on mobile devices and small-form factor screens are also available. One such video database for mobile devices is the LIVE Mobile Video Quality (VQA) database created by the researchers from The University of Texas at Austin (Moorthy, Choi, Bovik, & de Veciana, 2012). Contrary to other video databases created by tertiary institutes for designated purposes (Fliegel, 2014), this LIVE Mobile VQA database specifically focuses on low resolution video contents (far inferior to full HD) which are usually played on mobile devices with small-form factor screens. Full HD video contents are not used in this database. In fact, there is no video database that is specifically designed for mobile devices with full HD screen and better to the best knowledge of the author, although full HD screen or better is becoming the mainstream

(e.g. iPad Retina and Samsung Galaxy S4/S5). In order to fully exploit the hardware potential of high resolution screens found on many popular mobile devices and to discover how they affect end user experiences, a general-purpose full HD video database is required for this study. The IRCCyM/IVC 1080i database, for example, is a general-purpose full definition database containing interlaced scanned video sequences which can be used for any study that requires full HD video sequences.

There is no single database can perfectly fit into the requirement of assessing full HD progressive scanned video sequences on mobile devices. This is partially because VQAs were not commonly conducted on mobile devices and therefore, academics did not design any video database specifically for such purpose. This is also the case for the IRCCyN/IVC 1080i database as its contents are interlaced scanned. Interlaced scanning is the heritage of CRT (Cathode Ray Tube) display devices, is rapidly losing its favor since the introduction of LCD (Liquid Crystal Display). The latest encoders such as VP9 and H.265/AVC have abandoned the support of field coding natively as such technology is deemed to be obsolete (Sullivan et al., 2012) by the video industry. Modern display devices such as flat panel LCD TV and OLED (Organic Light-Emitting Diode) do not support the interlaced scanning mode natively (Lim, 1998). Although modern digital devices are capable of playing back interlaced scanned video contents by de-interlacing the content through internal software or hardware processing, artifacts are usually introduced in the process, causing deterioration of the video quality hence affecting the QoE (Quality of Experience) (Fan, Lin, Chiang, Tsao, & Kuo, 2008). As a result, the latest cameras from manufacturers are capable of producing progressive scanned video sequences natively; digital video contents have been produced and distributed in progressive scanning workflow recently as well.

The IRCCyN/IVC 1080i database best fulfils the experiment goals of this study subjectively assessing full HD video sequences on mobile devices with high resolution screens; providing ample variety of video contents. However, certain levels of modifications such as sub-sampling of colour space and de-interlacing must be made to the original sequences for VQA on mobile devices.

2.3.2 Encoder Bitrate Setting

For conducting VQA on mobile devices, the encoder bitrate configurations should be investigated because it has direct impact on both user perceived video quality and objective assessment scores. Therefore, suitable encoder upper and lower bitrate limits and the intervals of bitrate are usually carefully determined in previous studies.

All non-lossless video encoders will introduce distortions to the encoded video sequences when compressing original video signals in binary format. Such distortions are closely associated with the encoder bitrate configuration, the most important encoding parameter. Bitrate can be considered as the amount of information used to describe a given series of video images. The higher the bitrate, the more information a video sequence will contain and the closer it will assemble the original video sequence. As a result, user perceived quality would improve correspondingly with the increase of encoding bitrate (Menkovski, Oredope, Liotta, & Sánchez, 2009). Due to the constrains of the Internet bandwidth and its associated costs, the bitrate of video contents delivered over the Internet to mobile devices is limited. Most of the video content providers such as YouTube deliver their full HD video contents in a maximum bitrate of 6000kbps, based on our own calculation (video file size divided by its duration). Such a bit rate is deemed to be extremely high for streamed content by both content providers such as YouTube and ISPs (Internet Service Providers) as it would require at least 6Mbps Internet connection for smooth playback. Furthermore, mobile devices, as they are wirelessly connected, have limited data traffic available.

According to a pervious study, the most commonly used video encoder nowadays, H.264/AVC, will have significantly less improvement in quality once the bitrate is increased beyond 6000kbps (Pinson, Wolf, & Cermak, 2010), compared to the 4th generation of video encoders such as VP9 and H.265/AVC (Rao, 2013). When conducting video quality assessment studies, it would be reasonable to set the bitrate upper limit of the distorted video sequences to 6000kbps only, because their quality improvement is expected to stabilize lower than 6000kbps and it is around the limit of the internet connection speed in developed countries (Wilkinson, 2014).

In order to determine the lower threshold of the encoded bitrate in our study, we conducted a visual evaluation of a video sequence encoded by the H.264/AVC encoder in various bitrates to determine the lowest bitrate. At 500kbps, extremely obvious quality deterioration such as pixilation and colour distortion occurred. The observation was made on a 21.5-inch 2013 model iMac that comes with full HD display. It is also speculated that the user observed quality deterioration for lower bitrates on mobile devices is lower than on larger size displays due to the nature of the HSV. According to previous studies, at least 5 bitrates are usually tested to ensure the accuracy of VQA (J. Ohm et al., 2012; Pourazad et al., 2012).

2.3.3 Objective Video Quality Assessment

oVQA methods are computerized arithmetic used to measure the quality of videos automatically by detecting and comparing the amount of distortions resulting from encoding, recoding and transmission. oVQA compares encoded video sequence with the raw sequence (Hore & Ziou, 2010) to generate a score. Both original and compressed video sequences have to be prepared in the raw YUV format for objective comparison. In other words, encoded and distorted video sequences have to be decoded to YUV format for oVQA.

Researchers from The University of Texas at Austin (Wang, Sheikh, & Bovik, 2003), give three reasons for conducting oVQA:

1. Quality control mechanism can utilize oVQA for quality control purposes. For example, video content distributors can use oVQA to ensure the Quality of Experience by end users without over committing resources by using oVQA based adaptive streaming.

2. Gauge the performance of video encoding system and algorithms. If different options are available for a particular task, oVQA will be able to determine which option is the most advantageous.

3. Serves as a basis for optimizing the performance of video encoding system.

Finding an automated approach to quantify the resulting quality degradation will serve as a foundation to predict and determine the perceived video quality. oVQA methods are efficient means to quantify the amount of distortion automatically (Wang, Sheikh, et al., 2003). Although almost all commonly used video quality models are fast and cost effective, they have not taken the Human Visual System (HVS) and viewing conditions into account (Guo & Meng, 2006). This has resulted in poor correlation with user perceived video quality. However, there are HSV-based oVQA methods available (Eskicioglu & Fisher, 1995; Martens & Meesters, 1998; Rohaly et al., 2000).

The technologies behind oVQA methods can be classified into 5 types: Media-Layer (Gustafsson, Heikkila, & Pettersson, 2008), Packet-Layer, Bit-Stream Layer, Hybrid, and Parametric-Planning (Hands, Barriac, & Telecom, 2008). Media-Layer models only use raw video signal to compute the video quality without taking any other factors such as device type and network condition into consideration. Therefore, this type of objective measurement is

especially suitable for comparing the performance of different video encoders. On the contrary, packet-layer models use video-stream header information for video quality prediction without actually assessing the video content. Although this category of methods is easy and convenient to carry out, their accuracy is relatively low compared to other types of methods. In the parametric planning models, predicted video quality scores depend on quality planner parameters of network and therefore prior knowledge and details about the experiment environment are required (Anegekuh, Sun, & Ifeachor, 2013). Given the scope of our VQA, only Media-Layer models will be considered in this study as only video codec level comparison will be made. Table 6 presents a collection of commonly used oVQA methods for evaluating Media-Layer types of distortions. The table also illustrates the advantages and disadvantages of different oVQA methods.

Table 3. Commonly used oVQA methods

Abbreviations	Methods	Principles	Pros	Cons
SSIM(Wang, Bovik, Sheikh, & Simoncelli, 2004)	Structural Similarity	Detect the structural intactness of encoded sequence as human visual system is sensitive in spotting structural distortions	Cost effect and easy to implement. Taking HVS into consideration and therefore highly relevant to perceived video quality and QoE	Does not work well on specific types of distortions. Not originally designed for motion picture
MS-SSIM(Wang, Simoncelli, & Bovik, 2003)	Multi-Scale Structural SIMilarity	An extension to SSIM paradigm	Cost effect and easy to implement. Taking HVS and viewing condition into consideration.	Not originally designed for motion picture and not widely used. Tool is not easily available
MOVIE(Seshadrinathan & Bovik, 2009; Seshadrinathan & Bovik, 2010)	Motion-based Video Integrity Evaluation	-	Specifically designed for moving picture and accurate	Not widely used and the tool is not easily available
MSE	Mean Square Error	Detect and compare the signal difference in distorted video with the reference video	Cost effect and easy to implement. Widely used by researchers	Have little association with HVS and extremely insensitive to certain types of distortions.
PSNR	Peak-Signal-to-Noise Ratio	Same as MSE	Same as MSE	Same as MSE
VQM(Pinson & Wolf, 2004; Wolf & Pinson, 2007)	Video Quality Metric	Measure the perceptual difference to human beings	Accurate and widely used. Takes HVS into consideration	Take longer time to compute than PSNR and MSE

The ultimate goal of oVQA is to get its result as closely associated to the end user perceived quality as possible. Therefore, newer generations of oVQA methods take how HSV works into consideration. This philosophy has enabled the development of visual error sensitivity based algorithms used by many prevalent objective assessment methods such as

SSIM (Structural Similarity), MOVIE (Motion-based Video Integrity Evaluation) and VQM (Video Quality Metric). All these visual error sensitivity based algorithms attempt to emulate how HVS works and calculate the perceived quality. However, the HVS itself is overly complicated and it is not entirely understood by modern scientists, and therefore, it is argued that all the arithmetic based on it would not be robust since numbers of assumptions are made (Wang, Sheikh, et al., 2003). For instance, SSIM and MS-SSIM assume the HVS has adapted to extract structural information, comparing structural distortion of image instead of error (Wang, Sheikh, et al., 2003).

Unlike HVS associated oVQA methods, MSE and PSNR are the conventional classic oVQA methods that are most widely used. MSE and PSNR are very similar, as MSE is simply a function of PSNR. It is expected that they will have the same performance (Chikkerur, Sundaram, Reisslein, & Karam, 2011). Both MSE and PSNR are classic error-sensitivity based models that generate metric values by comparing the signal (data) and error (distortion) caused by encoder compression (Hore & Ziou, 2010; Huynh-Thu & Ghanbari, 2008). The more closely the encoded sequence resembles its original sequence, the higher the PSNR score that the test video sequence will receive.

PSNR and MSE are often criticized for their questionable association with the user-perceived video quality. Because PSNR treats any form of modification done to the original image as distortion or error, its result might not align with the user perceived quality under certain circumstances (Eskicioglu & Fisher, 1995; Girod, 1993; Teo & Heeger, 1994; Winkler, 1999). In some cases, enhanced images are more appealing to human observers, although the original image has been distorted to great extent (Savakis, Etz, & Loui, 2000) and therefore received low PSNR and MSE scores. As long as they are not enhanced to the extreme extent that causes them to lose naturalness, amendment made to original data remains as a means of enhancement (Yendrikhovski, Blommaert, & de Ridder, 1998). Existing studies suggest colourfulness and sharpness are more important factors than noise and pixilation, in terms of user perceived video quality. Therefore, PSNR can be inaccurate in some cases. Besides, PSNR underperforms in discriminating structural information of videos. When different types and levels of distortions are applied to a particular reference video sequence to test sequences, similar PSNR scores are produced for different test sequences under certain situations (Lu, Wang, Bovik, & Kouloheris, 2002). Further, PSNR and MSE models were not originally designed for oVQA; instead, they were designed for still picture testing decades ago.

Different from PSNR and MSE that do not take the HVS into consideration, some models assume human visual perception is extremely good in picking up structural information from video sequences (Wang et al., 2004). Such models include MOVIE, SSIM, SS-SSIM and VQM. These oVQA methods are full reference methods which require the presence of reference sequence. The studies conducted by the creators of SSIM (measures luminance, contract and structure), SS-SSIM and VQM, revealed that these methods outperform conventional methods such as MSE and PSNR (Wolf & Pinson, 2007). Corresponding to that, studies conducted by other researchers revealed the best performing oVQA models are SS-SSIM, VQM and MOVIE (Chikkerur et al., 2011).

oVQA methods can be categorized into three types based on how they function: full-reference, reduced-reference and no-reference. Most of the oVQA methods require having the original reference signal for comparison with the distorted ones, namely full-reference (FR) methods. However, undistorted video is commonly not available in the real world scenario as post production and compression for transmission will definitely apply certain level of distortion to it. Therefore, this gives rise to no-reference (NF) and reduced-reference (RR) methods that operate without (Wang, Bovik, & Evan, 2000; Wang, Sheikh, et al., 2003), and partial presence of the original signal respectively (Wang & Simoncelli, 2005). Since most RR and NR oVQA methods are video codec and transmission technology dependent (Farajzadeh & Mazloumi; Mu & Mauthe, 2008), they will not be considered for this study.

Each oVQA model is designed with slightly different philosophies and assumptions, they have their own advantage and disadvantages. In order to minimize the chance of getting unreliable data caused by the nature of different models, at least two assessment methods with different designs are often used for oVQA study.

There are readily available tools to carry out oVQA. Metrix MUX Visual Quality Assessment Package (Gaubatz) and MSU(Lab, 2013) are available. If only SSIM and PSNR scores are required, Video Quality Measurement Tool (VQMT) from Multimedia Signal Processing Group (MMSPG) (Hanhart, 2013) is available.

2.3.4 Subjective Video Quality Assessment

sVQA is a personal opinion based psychological test which involves human being as the participants (evaluators) who suppose to give their own opinions and judgments about the quality of the testing video sequences (NTT) after watching them. Since the end users of all video contents are always human beings, sVQA methods are generally regarded to be more

accurate than objective video evaluation methods and are frequently used as baselines to examine the performance or accuracy of oVQA methods (Seshadrinathan et al., 2010).

Although oVQA methods have recently earned the favour of many researchers due to its cost effectiveness, they are initially designed for still pictures and can become overly sensitive to certain types of distortions, producing untrustworthy figure (Wang et al., 2004). oVQA methods are said to be mechanical approaches which have little association with the HVS, and therefore its outcome is not significantly related to end user experiences. Some researchers argues that a single pixel shift might result a significant variation of score, which might not be perceivable by human beings even under careful examination (Korhonen & You, 2010). Therefore, only sVQA would be able to reveal the true video quality perceived by end users.

sVQA methods also have their limitations. Although sVQA methods are deemed by many researchers to be a reliable way to gauge the performance and accuracy of oVQA methods, they require researchers to commit a significant amount of resources (time, labor and financial cost) to perform. In real-life situations, the available resources are limited and important factors of how an experiment can be feasibly carried out. Additional, some sVQA methods have drawbacks caused by their design. For example, the ACR (Absolute Category Rating) method is known to have a memory effect, which means the current video rating is always affected by the quality of the previous video watched by the test subject (Hoßfeld et al., 2011).

To ensure the accuracy of scores gathered in sVQA studies, researchers usually take three main factors into consideration while designing their subjective studies (see Table 4).

Table 4. Factors Affect sVQA

Contributing factors	**Examples**
Environment of assessment	Viewing condition, ambiance light angle of display device, decoding hardware performance, decoding software performance etc.
Properties of tested video sequences	Content type, encoding codec, encoding parameter, frame rate, resolution bitrate, colour space etc.
Test participants	Psychological factors, eyesight, (anticipation), human visual system, personal preference, whether experienced or inexperienced tester etc.

When testing how the properties of tested video sequences affect the subjective user perceived video quality, the two factors, test subjects and environment of assessment, have to be kept constant while the properties of test video sequences are adjusted for different study purposes. Many of the previous studies followed the recommendations made by ITU (International Telecommunication Union) and VCEG (Video Coding Experts Group) about how the test environment for sVQA can be set up and the selection criteria for participants.

Such recommendations are based on the technologies available more than a decade ago and hence primarily devised for VQA studies on conventional display devices such as CRT TVs (ITU-R, 1998). Table 5 illustrates the related recommendations.

Table 5. Commonly used ITU recommendations for sVQA

ITU-T J.140 (1998)	Subjective picture quality for digital cable television system
ITU-R BT.500-13 (2012)	Subjective assessment of video quality for television picture
ITU-R BT. 710-4 (1998)	Subjective Assessment for image quality in high-definition television (HDTV)
ITU-T P.910 (2008)	Subjective video quality assessment for multimedia application
ITU-T P.911 (1998)	Audio-visual subjective quality assessment for multimedia application
ITU-T P.912 (2008)	Subjective video quality assessment methods for recognition tasks
ITU-R BT.1128-2 (1997)	Subjective assessment for image quality in standard definition digital television
ITU-R BT.1129-2 (1998)	Subjective assessment for image quality in standard definition digital television
ITU-R BT.1788 (2007)	Audio-visual subjective assessment of video quality in multimedia applications

According to the existing literature, there are a number of sVQA methods that are commonly used by scientists and researchers. The most prevalent ones are compared in Table 6 in terms of their principles, pros and cons.

Table 6. Commonly used sVQA methods

Abbreviations	Methods	Principles	Pros	Cons
ACR(ITU-T, 2008)	Absolute Category Rating	Single stimulus method, by which test subjects assess video clips for 10s and rate the video with a five-grade scale. Represented in MOS	Not time consuming, easy to setup, drafted and recommended by ITU-T(ITU-T, 2008). No Reference sequence required.	Sequence and memory effect.
ACR-HR(ITU-T, 2008)	Absolute Category Rating with Hidden Reference	Manipulation of ACR result by using a formula in order to prevent assessment results being affected by various types of content. Results represented in DMOS	Accurate, drafted and recommended by ITU-T(ITU-T, 2008) and used by established organization such as VQEG	Sequence and memory effect. Reference sequence required
DCR / DSIS(ITU-T, 2008)	Degradation Category Rating	Double stimulus impairment scale (DSIS) method, in which test participants assess reference video clips and distorted clips in pairs and rate the quality with a five grade scale. Results represented in DMOS	Adopted by EBU, accurate in telling minor quality differences	Might consume twice as much time as ACR and ACR-HR. Reference sequence required and hard to set up as dual screens are needed.
PC(ITU-T, 2008)	Pair Comparison	Test subjects are presented with a pair of videos, of which to determine video of better quality	Able to detect minor difference in quality	Sequence effect, Take longer than ACR or DCR, result is relative score
DSCQS(ITU R Rec, 2012a)	Double Stimulus Continuous Quality Scale	Test subjects are presented with a pair of videos, of which to determine the quality difference in a 5-grade scale continuously.	Largely used by tele-broadcasts, continuous assessment of the difference between two videos, result is absolute,	Hard to setup, test subject must be trained, time consuming, require reference video (FR)
SSCQE(ITU R Rec, 2012a)	Single Stimulus Continuous Quality Evaluation	Test subjects are presented with a stream of video which consists of various scenes. Quality is determined in a continuous fashion.	Reference video not required, less time consuming.	Various scene should be used, test subject must be very well trained. Participants might lose concentration is the test is too long. Reaction delay issues.
MLDS	Machine Learning Difference Scaling	Automated machine learning approach to determine video quality	Universal and not hardware dependant	Only assess bit rate factor, hard to set up. Reliability of the method is still under assessment. More oVQA oriented.

As revealed by Table 5, almost all sVQA methods are set forth by ITU. SSCQE, ACR and ACR-HR do not require the presence of reference sequence (Hidden Reference) are less resource consuming to carry out, whereas PC, SSCQE, MLDS and DCR will consume significantly more time although the results generated from which are likely to be more accurate. Unlike other methods, PC does not generate absolute results. Instead, the results generated from it are relative. ACR, ACR-HR and PC suffer from sequence effect adversely impacts the result accuracy. In term of rationale, ACR, ACR-HR and SSCQE are similar because they all aim to find the absolute score while DCR, PC and DSCQS are trying to find out the differential score by making comparison between two videos (ITUR Rec, 2012a). Such methods are not feasible for small factor display devices as there is a lack of means to synchronize the screen of two mobile devices. All the listed methods have different pros and cons; most of the listed methods are well tested and have details in the official documentation ITU-T recommendation P.910, BT-500.

In 2010, Prof. Antonio introduced a new method, MLDS (Menkovski, Exarchakos, & Liotta, 2010), based on a different concept from others. MLDS combines machine learning and method similar to PC to deliver sVQA scores. Since machine learning is involved, MLDS is difficult to carry out. Additionally, the reliability of the newly proposed method is not verified and there are not many researchers using it.

ACR is deemed to be the most commonly used subjective quality assessment methods by many academics (Song, Tjondronegoro, & Docherty, 2010) due to its cost effectiveness. Since ACR uses the 5/9/11-point scale as recommended by ITU-T (ITU-T, 2008), some researchers argued the scale is limited as it is difficult to map end users' perceptions onto a 5-point scale labeled as Excellent, Good, Fair, Poor and Bad (Sasse & Knoche, 2006).

In relation to the latest generation of video encoders such as VP9 and H.265/AVC, no sVQA study for VP9 (although oVQA was conducted for VP9 on large display devices (Řeřábek & Ebrahimi, 2014)) on mobile devices is ever conducted to the best knowledge of the author while sVQA for H.265/AVC on both large and small-form factor displays have been carried out by Garcia and Horowitz (Garcia & Kalva, 2013; Horowitz et al., 2012).

2.3.5 Correlation Metrics

Since both oVQA and sVQA methods have shortcomings, researchers often use both methods performance evaluation of video encoders. The correlation between the two sets of methods has to be established to demonstrate the reliability of data by using correlation

metrics. Assuming the correlation between sVQA and oVQA is linear, researchers usually use linear regression to illustrate the correlation: the higher the R value achieved, the stronger the correlation (Neter, Kutner, Nachtsheim, & Wasserman, 1996). The following commonly used metrics are adopted to evaluate the accuracy of oVQA methods:

Pearson Correlation Coefficient (PCC)

PCC is the commonly used term for the actual coined term Pearson Product Moment Correlation (PPMCC). This metric is commonly adopted to establish linear correlation between sets of data, with respect to how well they are related (Lee Rodgers & Nicewander, 1988). In the case of video quality assessment, sVQA and oVQA should produce high PCC values to validate their consistency to each other. The PCC value ranges from -1 to 1. If a strong positive linear correlation is discovered from the data samples, the PCC moves closer to 1; if the PCC value approaches to -1, it can be concluded that there is a strong negative linear correlation in the data samples. When the PCC value nears 0, no linear correlation discovered. In both cases that PCC values near -1 and 0, either sVQA or oVQA or both might have gone wrong. PCC is calculated as follows:

$$PCC = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Spearman Rank Order Correlation Coefficient (SROCC)

SROCC, the nonparametric version of PCC, measures the monotonicity of a metric (Ramsey, 1989). Similar to PCC, SROCC is the correlation coefficient between the predicted MOS and the subjective MOS scores. Non-linear correlations (e.g. parabola) will receive very low SROCC scores. The coefficient value of SROCC ranges from -1 to 1 and is calculated as follows:

$$SROCC = \frac{\sum(x_i - x')(y_i - y')}{\sqrt{\sum(x_i - \mathrm{y}')^2}\sqrt{\sum(y_i - y')^2}}$$

2.3.6 Performance Prediction Model

Due to the fact that H.265/AVC and VP9 encoders were released quite recently, researchers are still in the stage of concept testing, verification and performance analysis. There is little research work done pertaining to modeling the performance of these video encoders on small-form factor screens. To the best knowledge of the author, only QoE modeling of H.265/AVC content has been carried out so far (Anegekuh et al., 2013; Nightingale, Wang, Grecos, & Goma, 2013). However, there is more relevant research done

for H.264 encoder in term of performance modeling. According to previous study (Raake et al., 2008) conducted for H.264 encoder, in order to predict the subjective score, models have subjective scores such as SSIM or PSNR and a series of independent variables as predictors are usually considered by academics. Careful selection of these predictors is the key to construct effective models.

Predictors, required in video quality prediction modeling, can be categorized into two groups. The first group consists of the characteristics of the video contents, such as spatial resolution, temporal resolution, frame rate, and Quantization Parameter (QP); the other group consists of the different characteristics of different video encoders such as the largest supported MB size and MB number (Khan, Sun, & Ifeachor, 2012; Khan, Sun, Ifeachor, Fajardo, Liberal, et al., 2010; Khan, Sun, Ifeachor, Fajardo, & Liberal, 2010). Based on the existing literature, researchers usually select a combination of these contents to devise a precise model (Anegekuh et al., 2013) best suited to their research goal. Some of these models are limited because they are too focused on certain types of parameters and therefore overlook combined factors that will influence the prediction accuracy. For instance, researchers are too focused on identifying the missing MBs and its spatial influence, and completely disregard the influence of the video content type (Pinson & Wolf, 2004). This study is limited because there is convincing evidence illustrating that video content type has significant influence over both subjective and objective assessment outcomes, and is the second most important category of factors in creating video quality prediction models next to the capabilities of video encoders (Song, 2012). According to Song, various video contents have vastly different levels of image complexity and temporal information and therefore video content definition is crucial in creating a video quality prediction model. It is very necessary to include video content definitions as predictors of the model. Temporal resolution (TI) and spatial resolution (SI) are effective predictors as they clearly define the various characteristics of different video content according to the motion intensity and level of image detail. A detailed method of TI and SI value calculation is specified and recommended in ITU-T recommendation P.910 (ITU-T, 2008).

Apart from the commonly used predictors, some researchers have taken the uncommon approach of considering the type of video frame (I-frame, R-frame and P-frame) as the parameter for codec rate-distortion model (Nightingale et al., 2013). However, these studies are limited and contradict the way modern video encoders work. These studies have overlooked the fundamental mechanism of how video frames are encoded and compressed: as

noted, a given R-frame or P-frame consists of both intra and inter MBs; the ratio of both types of MB also varies significantly according to SI and TI. The type of video frame is not a strong predictor. Instead, it would be more logical to consider the total number of intra and inter MBs or their ratio as predictors of a video prediction model.

According to a study (Nightingale et al., 2013), predictors of video quality prediction models can also be considered as the combination of a few other predictors as long as there is correlation between them. For example, if there is correlation between spatial resolution and MB block size, the values derived by multiplying or dividing spatial resolution and MB values can be considered as one parameter of the model. If the designed prediction model cannot achieve the desirable prediction accuracy, researchers will usually adopt this option as the last resort.

Lastly, by using the sVQA scores as the dependent variables, the proposed model predictors are usually tested by using the stepwise linear regression function in data analysis software such as SPSS for their prediction accuracy (Song, 2012).

2.4 SUMMARY AND IMPLICATIONS

The review of the existing literature suggests that researchers have not yet carried out any subjective experiments on small-form factor screen devices to the best knowledge of the author. Although there are H.265/AVC sVQA studies carried out by researchers, the display device used are usually large full HD TV sets.

Different sVQA methods have different strengths and weaknesses. Although double stimulus sVQA methods are slightly more accurate, the extra time, labor, complexity to setup and cost do not justify the accuracy advantages. On the contrary, single stimulus methods such as ACR and ACR-HR are cost effective and easy to implement. Additionally, all the sVQA methods are provisioned by the standards set forth by ITU: the prescribed steps by ITU to carry out sVQA. Details such as number or group of desirable participants and sVQA environment set up are all included in the ITU documents.

Regarding oVQA, the available methods range from the widely used conventional ones designed since CRT TV era such as PSNR and MSE, to the newly emerging ones such SSIM and VQA which takes the HVS into consideration favor of many researchers in more recent years. Conventional oVQA methods such as PSNR do not take HVS into consideration. Which compare the error signal (noise) generated by distorted video sequence with the original one, such methods are prone to inaccuracy and error. On the contrary, newer methods

proposed by researchers, such as SSIM, UQI, MOVIE and VQM, make assumptions on certain characteristics of HSV, trying to align their outcome with the user perceived quality. These methods are proven to be superior to the conventional ones by existing studies. There are readily available tools to generate oVQA scores of different methods. A specifically designed tool provided by University of Texas will provide a database of videos oriented to oVQA studies on mobile platform. Software such as Metrix MUX Visual Quality Assessment Package (Gaubatz) and MSU (Lab, 2013) will be used to generate oVQA data.

Video databases are readily available for this study. However, because many of the databases do not provide the desirable resolutions and distorted bitrates, modifications of the databases have to be made for different research purposes. At least one encoder of the previous generation is required to produce distorted sequences for performance comparison with the latest generation of encoders.

Researchers commonly used the previous generation of video encoder, H.264, as a datum to assess the performance of the latest generation of video encoders. Due to the fact that the implementations of H.264/AVC standard are already matured, widely used and even commercialized for years, it would not be necessary to use the reference encoder of H.264 to produce the distorted sequences.

Both sVQA and oVQA outcomes should show consistency in order to validate themselves. If there is no consistency shown, researchers should consider revising their VQAs. Once data gathered from both sVQA and oVQA are in place, represented in the appropriate formats with proven statistical significance, PCC and SROCC can be calculated to illustrate the correlation or monotonicity between the sVQA method and the sVQA methods. After both sVQA and oVQA study outcomes are proven to be reliable by each other by correlation metrics, various model predictors can be tested for their prediction accuracy in the model.

Chapter 3: Research Design

The research goal has been determined as creating the user perceived video quality prediction models of H.265 /AVC and VP9 encoders on small-form factor screens, as stated in Chapter 1. In order to achieve this goal, this research is devised according to the research questions outlined in Chapter 1, along with the research gaps discovered in Chapter 2. This Chapter explains the research framework and discusses the methodologies adopted within the framework.

3.1 RESEARCH FRAMEWORK

This study has a 3-stage structure: exploration, experiment and analysis (see Figure 8). In the exploration stage, existing related studies and knowledge would be reviewed and analyzed according to their relevancy to the research questions. After a firm understanding of the prior knowledge is achieved, studies are then devised and carried out to gather the required data of sVQA and oVQA studies as the second stage. In the last stage, the collected data will be analyzed to select a range of suitable predictors to create perceived video quality prediction models that estimate the QoE on small-form factor screens for the 4th generation of video encoders. The details of the 3-stage research structure adopted for this study are illustrated by Figure 8.

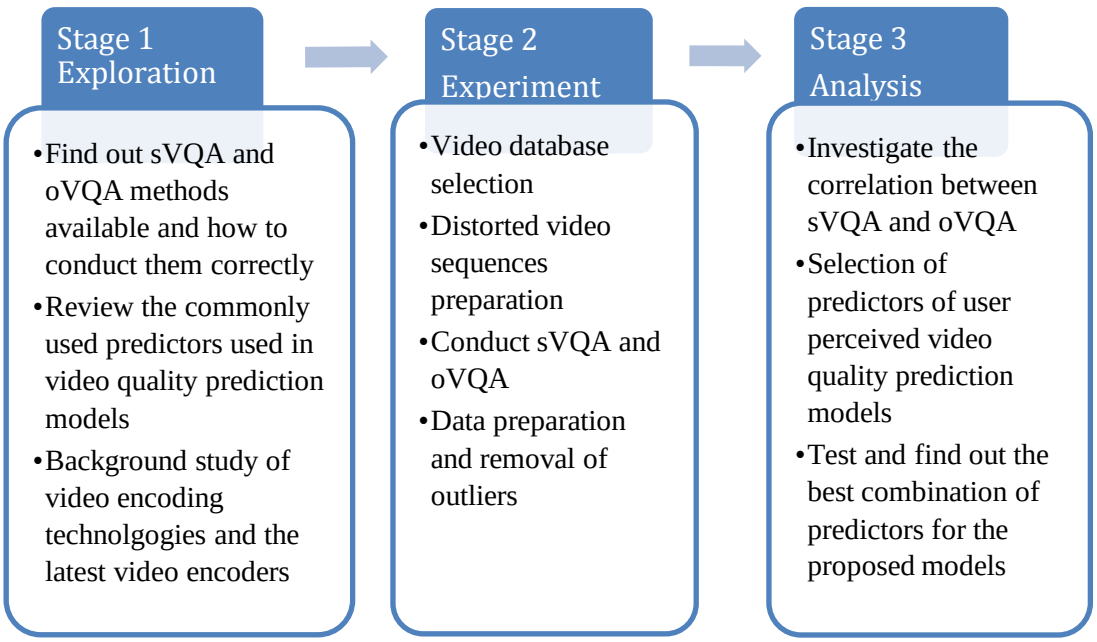


Figure 8. Research Framework

3.2 STUDY DESIGN

This study aims to build a video-quality prediction model to determine the user perceived video quality by using the data gathered from different types of video quality assessment methods. Figure 9 illustrates the workflow of the study.

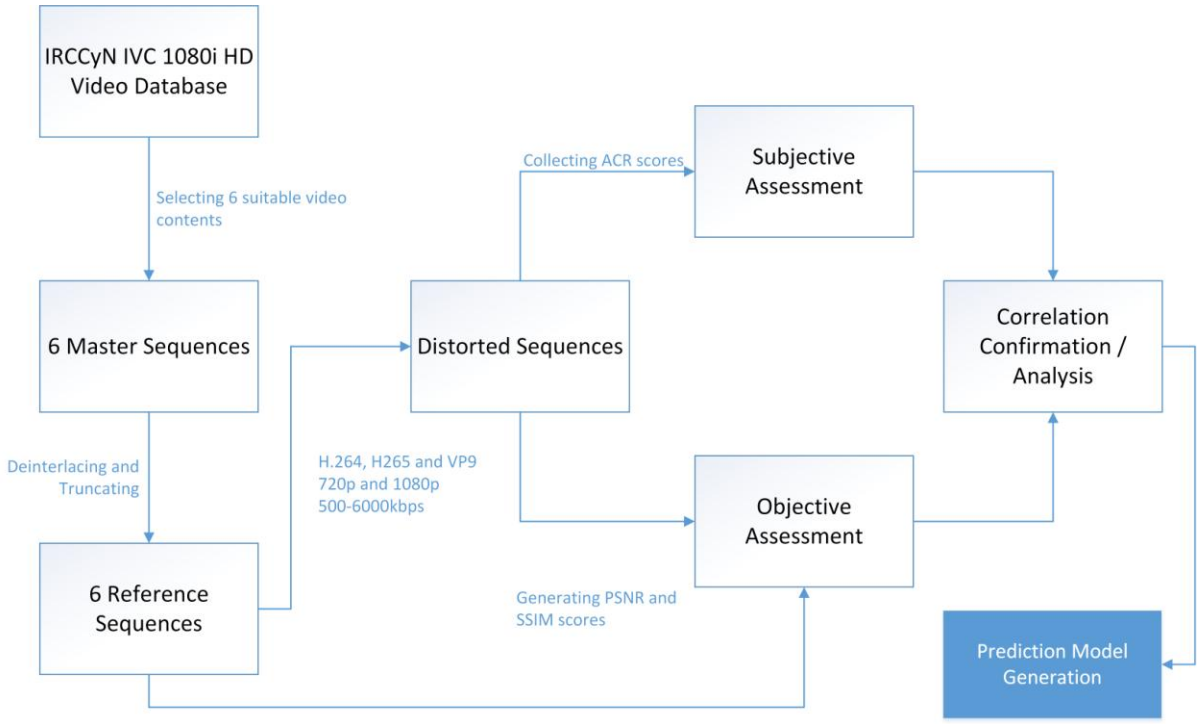


Figure 9. Study overview

First of all, a set of distorted video sequences is selected from the IRCCyN database and prepared to gather both subjective and objective data for predictor selection and model creation. Data gathered by different approaches can be categorized into sVQA and oVQA (Kreis, 2004; Sayood, 2002). Readily available sVQA and oVQA methods that best suit the purpose of this study are selected and carried out. The performance of VP9 and H.264/AVC encoders are accessed by both subjective and objective methods; due to technical constraint, the H.265/AVC encoder is accessed by only objective methods. ACR with a 9-point scale will be used to access the subjective quality perceived by 30 participants in this study. No reference video sequence will be shown to the participants during the sVQA session.

oVQA methods of PSNR and SSIM are to be used in this study. Based on the PSNR scores, the Bjøntegaard PSNR bitrate saving will be calculated to confirm and quantify the performance superiority of the 4th generation of video compression encoders, giving us more detailed reference to confirm the assumption that the 4th generation video encoders selected for this study are superior before conducting the resource intensive sVQA study to create the

performance prediction model. The Bjøntegaard model is used to illustrate the difference in percentage of bitrate saving of two rate-distortion (RD) curves and the average PSNR differences (Bjontegaard, 2008). As with sVQA, the available oVQA methods are separated into three categories: Full Reference (FR), Reduced Reference (RR) and No Reference (NR) respectively. Since it is widely known that NR and RR methods are feasible only when the types of distortions (e.g. transmission, wireless signal interference) are known and hence the encoding is arithmetic-dependent (Zhu, Asari, & Saupe, 2013), our oVQA study will be focusing on FR methods exclusively, in alignment with the research scopes stated in section 1.3 of Chapter 1. This study will not look into other factors such as network condition and bandwidth, that would affect the end user perceived video quality.

For carrying out both of these quality assessment methods in this study, reference sequences and distorted sequences are prepared. Some of the encoding parameters used to prepare the distorted video sequences will be used as the predictors of the proposed perceived video quality prediction model. The combination of factors of different test conditions includes content type, bitrate, resolution and codec. Distorted video sequences are produced by compressing the reference sequences (not the master sequences) by using three video encoders at different bitrates. Resolution scaling, video resizing and de-interlacing of master sequences to product reference videos are done by using FFmpeg exclusively. Pixel format (YUV), frame rate and scanning mode were standardized in this study and therefore will not be considered as the predictors that will affect the user perceived video quality.

Scores from sVQA and oVQA are then analyzed for the selection of predictors and creation of the models.

3.2.1 sVQA Design

sVQA methods share the same goal as the prediction models: estimating the user-perceived video quality as accurately as possible. Therefore, the subjective score gathered from the sVQA study will be the independent variable representing the user-perceived video quality for creating the prediction models.

The sVQA study is designed to simulate the real-life situation as closely as possible in the laboratory environment while maintaining a certain level of control that will ensure the efficiency of the study and the accuracy of the collected data. However, small-form factor devices are usually mobile and are used under various conditions such as at home or on the go. We therefore take various measures to ease the stress of participants, including recreating a

user's environment rather than a laboratory setup. We also divided each sVQA session into a few sub-sections and the participants were allowed to take a break between sub-sections. Each sub-session lasts approximately six minutes, the usual length of online video content on websites such as YouTube. Typical short video contents of common scenarios are used. Psychological factors of participants that might influence the study outcome are also kept at a minimum level. For example, the facilitators do not sit next to the participants to ease their psychological pressure while they are assessing the distorted video sequences.

The ACR method recommended by ITU-T (ITU-T, 2008) was used for sVQA. ACR is often criticized for its commonly used 5-level scale, Excellent, Good, Fair, Poor and Bad (Sasse & Knoche, 2006). Simply mapping participants' perception onto a 5-level scale is not seen as sufficient. Therefore, instead of the classical 5-level scale adopted by most sVQA studies, a 9-level scale for ACR adopted for this study keeps the five recommended labels and adds 4 additional levels. The ACR scores are expressed in Mean Opinion Scores (MOS), where the votes by the 30 participants are illustrated.

Additionally, we recruited our participants by using a set of criteria to make sure they are free from visual associated diseases and from expected user groups. Participants were placed in a controlled lab environment for the sVQA.

The MOS is calculated as:

$$MOS_i = \frac{\sum_{j=1}^{N} s_{ij}}{N}$$

Where N represents the number of valid participants and s_{ij} is the score by participant j for the test condition i. The connection between the estimated mean values from a collection of participants (30 in total) and the real mean values of the participants is given by the confidence interval (CI) of the estimated mean. The $(100 - 100\alpha)\%$ CI for MOS is calculated by using the student's *t*-distribution:

$$CI_i = t(1 - \frac{\alpha}{2}, N) \cdot \frac{\sigma_i}{\sqrt{N}}$$

The statistical significance between two MOS values generated by VP9 and H.264/AVC encoders are verified by a *t*-Test. The *T*-test is conducted because the sVQA score sample size of this study is relatively small, with only two variables compared. To compare more than two variables, Analysis of Variance (ANOVA) is commonly used.

3.2.2 oVQA Design

PSNR and SSIM are used to generate objective scores in this study. The two methods are commonly used by academics and researchers from video industry, and do not consume significant amounts of resources to compute. Although the literature review revealed that PSNR is inaccurate under rare circumstances as it does not take HVS into consideration, it is still commonly used for benchmarking purposes by researchers. We therefore selected the classic PSNR as one of our oVQA methods. Additionally, the PSNR score will be used for calculating the bitrate saving by using Bjøntegaard model.

The other oVQA method selected is SSIM, which takes HVS into consideration by assuming HSV is extremely good in extract structure information. The two oVQA methods selected will complement each other to ensure that the accuracy of the data collected for oVQA as a whole is not severely affected by the limitations implied by the difference in natures of the two different methods. In case that inconsistency between PSNR and SSIM scores is discovered, the oVQA will be reviewed and additional oVQA methods will be introduced.

As noted in section 3.6, although sVQA methods are deemed to be more accurate and are usually used as a datum to examine the accuracy of oVQA methods, they have some memory effect. The oVQA is therefore also used in this study to confirm the accuracy of the sVQA method.

3.2.3 Video Materials Preparation

We selected six master video sequences from the IRCCyN (Institut de Recherche en Cybernétique de Nantes)1080i HD Video Quality Database (Péchard, Pépion, & Le Callet, 2008), which is freely available to researchers and academics for research purposes, to prepare the reference sequences. The video database consists of both distorted and undistorted Full HD videos in uncompressed YUV4:2:2 interlaced scanning format at 50 fields per second without audio. The available distorted video sequences contained in the database were not produced according to the requirements of this study and therefore were unsuitable to be used. We have to generate our own distorted sequences. The six selected master sequences were de-interlaced and sub-sampled to produce both reference and distorted sequences. Two additional master training sequences were selected to create 2 sets of distorted videos for our training session. **Distorted sequences were tested against the reference sequences instead of against the master sequences in the objective assessments**.

Since most video contents online are progressive scanned 4:2:0 sampled sources, to generate the reference sequences, we de-interlaced and sub-sampled the six master sequences from YUV4:2:2 50i to YUV4:2:0 25 frames per second in progressive scanning format by using FFmpeg (Bellard & Niedermayer, 2012). FFmpeg was also used for up-scaling and down-scaling the 720p resolution video sequence. Although all three video encoders used in this study are capable of sub-sampling, the third party application, FFmpeg, is used to pre-subsample (down-scale) the master sequences to 720p resolution in order to prevent inconsistency that might be caused by potential difference in the resizing arithmetic of different video encoders before encoding. After the distorted or compressed video sequences are decoded by their respective video decoders, the distorted raw YUV file in 720p resolution will again be up-scaled by using FFmpeg for consistency measures.

As noted in Chapter 2, the HVS is not as sensitive picking up colour information as picking up textual information and therefore it is unlikely that end users will be able to tell the difference between YUV4:2:0 and YUV4:2:2 sequences. Evidence from existing literature suggests identical notion (Sullivan, Topiwala, & Luthra, 2004). Moreover, all online video content distributors such as YouTube and Dailymotion deliver their videos in the YUV4:2:0 sampling rate. It is unlikely that any end user will come across videos of higher or lower sampling-rate under normal circumstances. Master sequences are de-interlaced in this study because less and less video contents are delivered in heritage interlaced scanning format to end-users as modern flat panel display devices display progressive scanning source natively. Another reason for choosing the progressive scanning mode is that the latest generation of video encoders such as VP9 and H.265/AVC encoders has abandoned the support of field coding. The latest generation of encoders are designed to encode progressive scanning video sequences only; interlaced scanning is considered to be obsolete and no longer used for display devices and content distribution (Sullivan et al., 2012).

The 6 master sequences last for 10 or more than 10 seconds. For standardisation, we truncate all six sequences to 10 seconds or 250 frames exactly, to produce our reference sequences. The six sequences consist of different types of scenes simulating real-life situations relevant to the video consumption habit of end users. Table 7 gives the detailed description of the selected video sequences; Figure 10 shows their screenshots. These scenes include fast motion (e.g., Seq 1, Seq 4), extreme portrait close-up (e.g., Seq 2), stage performance under a poorly lit condition (e.g., Seq 3), a long shot for sports games with

panning camera (e.g., Seq 5), and high image details and extreme colour contrast contents (e.g., Seq 6).

Figure 10. Screenshots of the selected video sequences

Table 7. Descriptions of Selected Video Sequences

No.	Scene	Description
Seq. 1	Fast moving object with high stationary spatial details in the back ground	Huge crowd of marathon runners moving from right to left with no panning
Seq. 2	Close-up portrait	Costumed ship captain looking through his monocular under heavy downpour with slow and minor zoom-in. Low contrast.
Seq. 3	Stage performance under lowlight condition	Singer walking across stage under spotlight with band members behind him. Random panning and zooming.
Seq. 4	Fast random motion	Costumed actors and actresses moving about randomly in the foreground and background
Seq. 5	Panning camera with fast motion	Wide angle soccer game
Seq. 6	Random motion with high contrast image	Costumed actors and actresses moving around in a park

Referring to Table 7, Master video sequence 1 is chosen based because it consists of extreme details and fast moving objects that put the ability such as inter and intra prediction of video encoders to test. Master video sequence 2 contains a close-up portrait that is commonly seen in television drama series, talk shows, news broadcast and movies where protagonists' facial expressions draw the attention of end users and play a major role in telling a story. In such scenes, the background of the video sequence is usually inter-coded with little motion transformation applied to the MBs while the moving object is intra coded. Master video sequence 3 was selected to demonstrate the situation where the viewer is watching the

recordings created under poorly and unprofessionally lit circumstances, such as candid shot, live music concert, and videos created in uncontrolled environment. Such scenes often contain dark background and gradients which are hard for encoders to encode. Master video sequence 5 was chosen as large groups of users often watch fast-paced sports games such as soccer, rugby and basketball where cameras are constantly panning with the fast-moving players. Due to the constant panning motion of camera, it is expected such video sequences will contain a large number of inter coded MBs with motion vectors. Master video sequence 4 and 6 were selected to simulate video contents consisting of high colour contrast. Such contents are often seen in movies as they have been processed extensively by postproduction techniques such as colour correction to create the mood and atmosphere directors want to convey to the audience or even to favor the characteristics of HVS. Such content will also put the ability of the encoders to compress colour information correctly to test.

The six reference sequences are then encoded into distorted sequences by using combinations of the following encoding settings:

- **Bitrate (BR): 500kbps, 1Mbps, 2Mbps, 4Mbps, 6Mbps**
- **Resolution (RS): 720p and 1080p**
- **Coding Format (CF): H.264, H.265 and VP9**
- **Content Type (CT): Sequence 1, Sequence 2, Sequence 3, Sequence 4, Sequence 5 and Sequence 6**

Since the three video encoders we used in this study do not compress the video signal by using lossless arithmetic, all of them will introduce distortions to the compressed video sequences. The level of distortions is closely associated with Bitrate – the most important encoding parameter. Bitrate can be considered as the amount of information used to describe a given image. In this study, bitrate varies between 500kbps to 6000kbps. Five different bitrates are created. A total number of 180 test sequences are generated from the six reference videos. Since content is denoted as “CT”, the total number of distorted sequences to be generated can be calculated by using the combination formula where *n* represents the total amount of choices available while *r* represents how many choices are made at a time:

$$\binom{n}{r} = nC_r = \frac{n!}{(n-k)!\,k!}$$

Table 8 illustrates the parameters we have control of and the number of choices / settings for each parameter. The multiplication of choices of BR, RS, CF and CT will give us the total number of distorted video sequences to be created in this study. 180 combinations are derived. In other words, 180 distorted video sequences are to be generated from the 6 reference video sequences.

Table 8. Number of distorted sequences

Parameter	**No. of Choice**	**Description**
Bitrate (BR)	5	500kbps, 1000kbps, 2000kbps, 4000kbps, 6000kbps
Resolution (RS)	2	720p25, 1080p25
Coding Format (CF)	3	H.264, H.265, VP9
Content (CT)	6	Sequence 1 to 6

Therefore, the total number of distorted sequences generated in this study can be calculated as such:

$$\mathbf{5 \times 2 \times 3 \times 6 = 180}\ \boldsymbol{combinations}$$

We adopted the Variable Bitrate (VBR) setting as our bitrate control mode to encode the distorted sequences. Unlike the Constant Bitrate (CBR), which uses the same amount of data (bits) to interpret every frame of a given sequence regardless of the complexity of the scene, VBR is able to adjust the amount of data used to represent each frame based on its complexity. Most of the online videos are encoded in VBR mode. As a result, video content encoded with a VBR setting is likely to be superior to those encoded by CBR if the overall file sizes remain the same. However, allowing bitrates to vary significantly in a given sequence will affect the overall size of the encoded video. We therefore allow marginal fluctuation of bitrates and make sure the final file sizes of distorted sequences correspond to their bitrates. For instance, a 10-seconds distorted sequence produced from a reference sequence with a bitrate setting of 500kbps (kilobit per second) is expected to have a file size of 610.3kB (kilobyte) approximately. Distorted sequences produced that do not meet the required sizes are rejected. The calculation is illustration by the following formula.

$$\frac{\textbf{Input bitrate}(\textbf{bit per second}) \times \mathbf{10}}{\mathbf{8192}} = \textbf{Expected File Size} \pm \mathbf{5}\%(\textbf{in kilobyte})$$

Table 9 illustrates the estimated file size of encoded video sequences for 5 different bitrates.

Table 9. Estimated encoded video file size

Bitrates	**Estimated file size (±5%)**
500kbps	610.3kB
1000kbps	1.22MB
2000kbps	2.44MB
4000kbps	4.88MB
6000kbps	7.32MB

We allow a tolerance of ±5% on expected final size of distorted video sequences.

3.2.4 Encoder Settings

The 3 video encoders used in the study are different in their designs. We adjusted the main encoding parameters to make the study a fair comparison. The configuration options used for all three encoders are displayed in Table 10.

Table 10. Encoder configuration

Encoder	**x.264**	**VP9**	**H.265**
Versions	FFmpeg version N-60329-ge708424	v1.2.0-3909-g8b05d6a	HM14 (Apr 2014)
Encoding value	Very slow	Max	-
Pass	1	1	1
Maximum MB Size	16x16	64x64	64x64
Bitrate Control Mode	VBR (defined target bitrate)	VBR (defined target bitrate)	VBR (defined target bitrate)
GOP Length (Intra Period)	320	320	320
GOP Size	Auto	Auto	8

We specified the GOP size of H.265/AVC to be the default value of 8 as it is a compulsory option to be turned on for the encoder. For HM14 H.265/AVC test model encoder, a configuration profile file was loaded. The profile is set to main profile fast search and all options are set to default except bitrate. Since the number of B-frame and P-frame is impossible to be kept constant for all three encoders due to design differences, we specified only the maximum I-frame distance for all three encoders. The I-frame is usually introduced at scene change or the beginning of a video sequence. However, each of our video sequences contains only one scene and thus probably only one I-frame is encoded in each of the distorted sequence.

3.2.5 sVQA Test Equipment

Video sequences were presented on Microsoft Surface Pro 1st Generation, which comes with a 10.6 inch screen of native resolution of 1920 × 1080 pixels. We installed the hardware with a fresh copy of the latest Windows 8.1 Pro. The operating system was then left to be updated automatically with the latest patches. The only third party software installed onto the device was Google Chrome and CCCP (Combined Community Codec Pack). Our web-based sVQA application and the test video sequences were copied onto the built-in solid-state storage device of the tablet for smooth playback. We made sure our device was capable of decoding both VP9 and H.264 sequences by using the Windows Media Player Basic that comes with CCCP to detect if there were any skip frames. Based on the real time statistics provided by the player while playing all the prepared video sequences, no skip frame was detected and there was no visual latency observed. By playing test sequences in both window and full-screen mode in our web application, we made sure there is no cropping or dark border caused by over-scanning and under-scanning. The hardware specifications of the device used in the sVQA study can be found in Table 11. The hardware specification of this device is mainstream in the year this study was conducted (2014).

Table 11. Mobile Device Specifications

Processor	3rd Gen Intel Core i5-3317U @ 1700MHz Dual-Core
Screen Size	10.6 inches
Screen Type	LCD
Resolution	1920x1080 pixels
Graphic Card	On-die Intel HD4000, Frequency: 650MHz – 1150MHz
Storage	64GB Solid-State Drive
Memory	4096MB
Operating System	Windows 8.1 Pro

Although the device we used has a touch screen for user input, a wireless mouse was paired with the device for fuss-free operation and fast training of the participants. The device was disconnected from the Internet and set to flight mode during assessment sessions. During the assessment process, the participants were left in the space alone to complete the tasks while we kept a distance away for observation without any form of disruption as participants may tend to be tense and indecisive while taking assessments in the presence of researchers (Song et al., 2010).

3.2.6 sVQA Test Environment

The subjective study was conducted at a controlled environment at Mobile Innovation Lab, Queensland University of Technology as recommended by ITU-R (ITU-R BT500) for home environment. Based on previous related experiments and suggestions made by researchers and ITU documents, a group of 20 to 40 test subjects is the minimum requirement for subjective studies (Song, 2012). In this study, all the participants were asked to sit in an enclosed space where there was no strong window light and any source of distraction. The table was set to normal height and the chair was adjusted to comfortable position for the test. Only florescent lights on the ceiling were used as the ambience lighting. For every subjective video assessment session, the luminance level in the space was set to constant and the brightness of the device screen was adjusted to automatic. We made sure there was no reflection of the light sources as bright spots shown on our device screen. The details of the setup can be found in Table 12:

Table 12. Specification of sVQA environment

Maximum Display brightness	350cd/m2
Maximum observation angle relative to the normal	30 degree
Screen size	10.6 inch
Screen resolution	1920×1080
Viewing distance	80cm

3.2.7 Participants

The subjective study conducted for participants is separated into two parts: training and actual evaluation. We preselected the participants by using a set of criteria and those who are found to be unsuitable to be removed from the test. Additional participants are recruited to make up the total required number of 30 participants. The demographic characteristics are as followed:

- Age
- Experience with video quality assessment study
- Visual impairment and visual associated illness (colour blindness and myopic etc.)

We recruited a total number of 30 participants aged between 20 to 35 from university students and staff mainly, regardless of gender. Participants who have video related background or experience were removed from this study. This arrangement was carried out to meet the requirement of 24 minimum subjects recommended by Video Quality Experts Group for sVQA (Hands & Brunnstrom, 2007; Péchard et al., 2008) and the ITU standard (ITUR Rec, 2012a). We made sure all our recruited participants do not have visual associated illness and impairment by asking them the details of the training video sequences while going through them. During the training session, participants were asked to describe the details, such as the colors of the objects shown in the training videos, and to make simple quality comparisons. Participants who were unable to fulfill these simple tasks were deemed to be unsuitable to sit for our study and therefore are removed from the study. No data were collected in such cases.

In order to get the participants familiarized with the operation of our sVQA program and hardware under the instruction and guidance of our facilitators, our facilitator explained in details about how our application and hardware work to each of the participants before they got started with the 2 sets of training sequences during demo sessions. During the demo session, the participants were free to get familiar with the evaluation interface and no time limited is enforced. Participants were allowed to try the demo session as many times as they desired until they were confident with the operation. After the demo session, participants could choose to start the sVOA study at anytime. Participants were informed that during the actual sVQA session, they were allowed to stop the assessment session anytime they intended should they felt uncomfortable or tired. Their scores, under such circumstances, would not be recorded.

3.2.8 sVQA Voting

The sVQA followed the ACR (ITUT Rec, 2008), whereby a subject viewed the distorted sequences one at a time and rated them on a scale after viewing each sequence. A limited voting time of 10 seconds was applied and no score was recorded in the case where the participant fails to make a selection. The voting process can be illustrated by Figure 11:

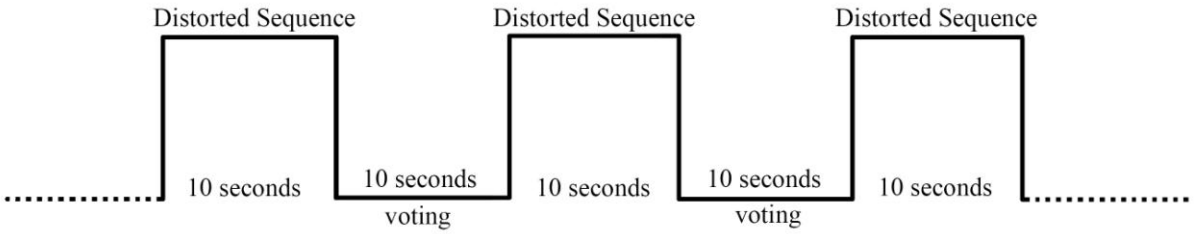


Figure 11. ACR voting interval

Since this is a No Reference (NR) study, no original video sequence was shown. To alleviate participants' fatigue, we divided the 180 distorted sequences produced from 6 reference sequences into six groups (i.e., mixing the bitrates, resolutions and codec) to be watched in six sessions. The participants are allowed to take a short break (about 5 to 10 minutes) between sessions. We adopted a 9-point quality rating scale, which is labelled as 9-excellent, 7-good, 5-fair, 3-poor, and 1-bad to indicate the subjects' opinions. Since Google Chrome is capable for the playback of both VP9 and H.264 sequences via HTML5 (Protalinski, 2013), we programmed a web-based application to play the distorted sequences back in full screen. The distorted sequences were arranged in a random manner regardless to their bitrate, resolution and encoded codec. However, the presentation order of test sequences for each sets of content is identical regardless of participants. Figure 12 illustrates the web-based sVQA application interface.

sVQA sessions from 1 to 6 and demo are hyperlinked in the homepage. Participants are able to move the mouse cursor and click onto the hyperlinks to go into each sVQA session. The order of the sVQA sessions is not enforced and it is observed most participants viewed the sessions in the order of 1 to 6.

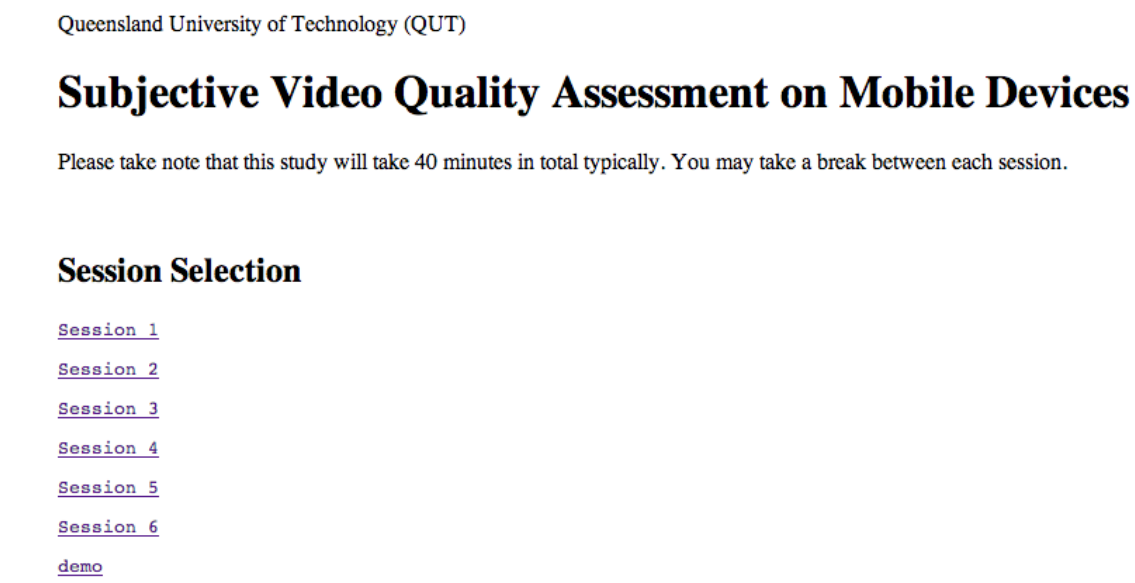


Figure 12. Index page of the HTML5 based application for sVQA (VP9 and H.264/AVC only)

During the sVQA session, the application requires participants to rate the watched sequences with a dialogue screen after each video sequence is finished. A sliding-bar will be displayed to the participant to rate the watched sequence on a 9-point quality scale. Upon clicking the "Next" (confirmation) button displayed below the sliding bar, participants' ratings will be recorded automatically in CSV files. Since participants have only 10 seconds to rate each video sequence, there is a countdown timer displayed on the top right hand corner of the rating screen to prompt the participants about the remaining time. If a participant misses the rating, the score for that video will not be recorded and therefore will not be counted towards the average ACR score. Since participants in this study have been given a short training period during the demo session, the chance of failing to vote before the timer goes down to zero is about 1/1000. The entire sVQA process is automated without the intervention of researchers. Unless participants call for assistance, we keep a comfortable distance away from them.

Figure 13. ACR 9-level voting scale page

Figure 13 shows the voting window that will pop up after each distorted video sequence finishes playing. The classic labels from Bad to Excellent remain unchanged while a 9-point scale is used. The background of the pop up voting window is set to grey to minimize distractions.

3.3 ANALYSIS TOOLS

The objective video quality assessment tool used in the study is Video Quality Measurement Tool (VQMT) from the Multimedia Signal Processing Group (MMSPG) (Hanhart, 2013). The tool is a Windows command line based application that generates both frame-by-frame and average PSNR or SSIM scores. Both metrics are implemented in OpenCV (C++) based on the original Matlab implementations provided by their developers. Distorted sequences were paired up with their corresponding reference sequences to generate PSNR and SSIM scores. Individual CSV files were generated for each pair of sequences and the statistical data contained in the files were stored tabulated in the Statistical Package for Social Sciences (SPSS) for Analysis.

In order to find out the PSNR based bitrate saving of each tested video codec, Matlab was used to calculate the Bjøntegaard metric (Bjontegaard, 2008) based on the PSNR scores retrieved by VQMT. The Matlab source code is freely available to all researchers (http://www.mathworks.com.au/matlabcentral/fileexchange/27798-bjontegaard-metric).

IBM SSPS Version 22 was used to tabulate and analyse the retrieved subjective and objective scores. The scores generated by sVQA contain outliers that affect the consistency of the dataset. The tool is also used for data outlier identification, correlation and difference analysis for sVQA and oVQA scores, discovering potential predictors and regression testing for prediction models generated. More details are given in Chapter 4 and 5.

3.4 RESEARCH DATA PROCESSING

No normalization is required for the SSIM, PSNR and ACR scores collected in this study. Processing of raw ACR scores is carried out according to ITU-R recommendation (ITUR Rec, 2012a).

oVQA scores do not contain outliers as they are generated automatically by computer. On the contrary sVQA scores will contain outliers where values significantly deviate from the

normal data collected. In this study, the ITU guideline on subjective score outlier removal is enforced (ITUR Rec, 2012a). 95% of normal data distribution is generally considered. However, if the sVQA scores garnered from this study have extremely uneven distribution, 90% can also be taken into consideration.

3.5 ETHICS AND LIMITATIONS

QUT official ethical clearance was obtained before the subjective study was conducted. No identifiable personal information of the participants was gathered. The digital data was stored in the hardware at QUT Mobile Innovation Lab with password protection.

sVQA methods such as ACR have memory (sequential) effect (Hoßfeld et al., 2011) whereby the perceived quality of a given video sequence is always affected by the previous one. For instance, a given distorted video sequence is likely to receive a higher score if the previous sequence has extremely inferior perceived quality. Although this sequential effect caused by ACR can be negated by rearranging the presentation order of the tested distorted sequence, it is not feasible in this study due to the constrain on the availability of resources (time and labour). The HTML5 based application will store the ACR scores in individual CSV files. To be specific, extracting these data and porting them into SPSS can be extremely resource consuming. Therefore, oVQA in this study has a role to verify if the data obtained from sVQA is affected by memory effect. However, such a measure will only be able to detect significant disparity caused by memory effect, if there is any. Insignificant level of memory effect is hard to be detected and quantified.

Due to the fact that the sVQA study has to be conducted under a controlled environment, we cannot recreate the real life situation of watching a video on a mobile device with high accuracy. For example, almost all video sequence online have multiple scenes, as noted in Chapter 2, and therefore it is logical to consider the performance of video encoders to be highly associated with the complexity of video image (image details) and the intensity of motion (speed and number of moving objects). However, in our sVQA study, we are unable to quantify the properties of the content of the 6 videos or to use them to further improve the accuracy of the prediction model.

Although content definition and characteristics of video encoders are taken into consideration as the predictors in the subjective video quality prediction model, it is not feasible to take the screen size as a predictor in the proposed model due to project time and cost limitation. In order to take mobile device screen size or resolution into consideration for

prediction model creation, the same video contents also have to be tested on different devices in sVQA too. This will potentially extend the subjective study duration for each participant by 3 to 5 times.

Chapter 4: Results & Analysis

For each pair of reference and distorted sequence, we averaged the PSNR or SSIM scores for the total 250 frames to generate the final PSNR or SSIM score. Since 180 distorted sequences were generated in total, 180 sets of PSNR and SSIM scores were collected and 3600 individual ACR scores were recorded for 120 distorted video sequences across 30 participants. Only ACR scores of VP9 and H.264/AVC encoded sequences are collected due to the technical constrain that H.265/AVC sequences cannot be played by our web-base sVQA application.

In this Chapter, the subjective and objective research data collected will be discussed in details. Based on the scores collected, both subjective and objective scores show high levels of consistency. Two objective metrics, SSIM and PSNR, showed consistent results.

4.1 SVQA OUTLIER REMOVAL

Before statistical measures were calculated to depict the sVQA scores distributed across the 30 participants for different test conditions, such as various combinations of content, codec, resolution and bitrate, outliers of the sVQA scores were removed according to the specifications of ITU recommendation (ITUR Rec, 2012a). According to the document, all mean scores are supposed to have an associated confidence interval derived from the standard deviation and size of the sample when the score of a test is presented. Although a recommend confidence interval of 95% is recommend, we found the ACR scores gathered from the study are quite sporadic and therefore a narrower confidence interval of 90% was applied in our study. In this study, 3597 ACR scores are recorded, as participants failed three times to give a score to a particular distorted video sequence within the given voting time for three times. 136 entries out of 3597 were hence removed from the ACR score list as outliers. The following formula is used to calculate the z-scores in SPSS. The value of 1.645 stands for the 90% confidence interval applied in this study.

$$\delta_{jkr} = 1.645 \frac{s_{jkr}}{\sqrt{N}}$$

4.2 SUBJECTIVE ASSESSMENT

From the study process illustrated by Figure 9, distorted videos produced from the 6 reference videos were rated by participants by using the ACR method. Figure 14 illustrates the average ACR scores of two video resolutions from 30 participants. Based on the figure, it is obvious that VP9 encoder performed significantly better than H.264/AVC encoder across all five bitrates. Both encoders demonstrated similar performance patterns when bitrates increased. The ACR scores for both encoders also progressed in a non-linear fashion. The correlation of bitrates and ACR scores is logistic or logarithmic in Figure 14. The wide error bars shown resulted from the different mean scores for the 6 different video contents. Table 13 lists the percentages of the average ACR score differences throughout all five bitrates by comparing VP9 to the H.264/AVC encoders. The t-test results show the significance of the ACR score differences. Only VP9 and the H.264/AVC encoders were tested by the sVQA due to technical constrains.

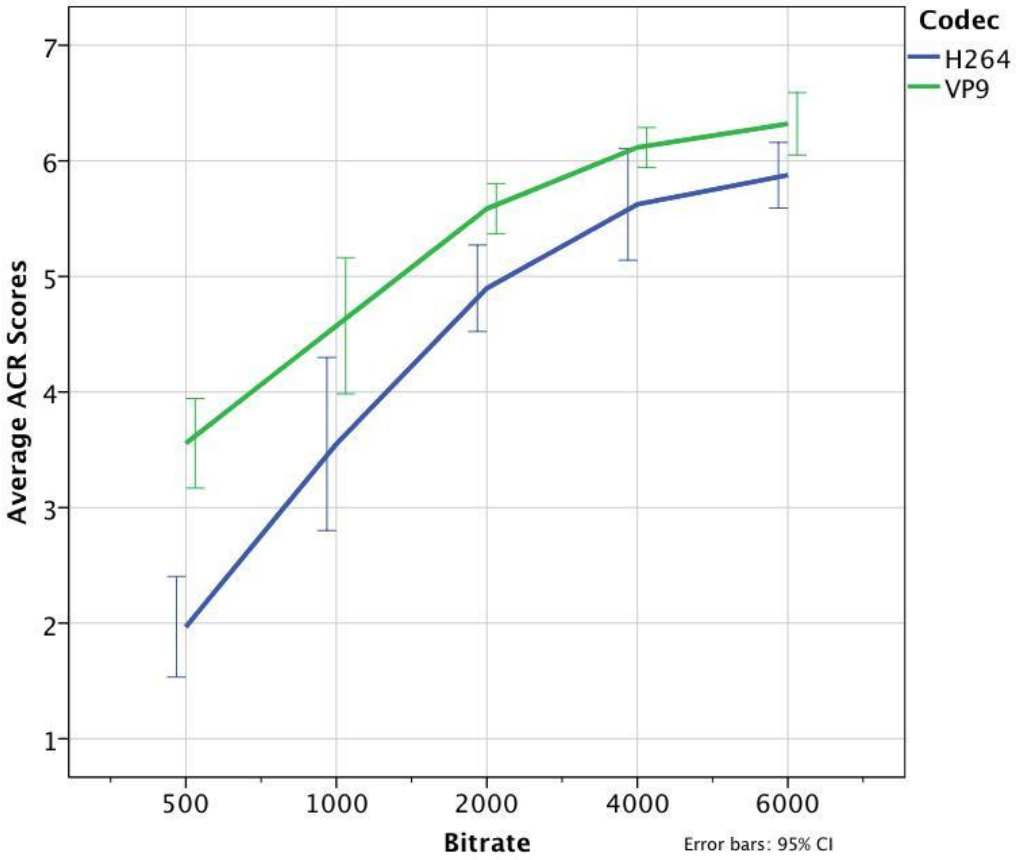


Figure 14. Average ACR scores

Based on Figure 15 and the t-Test results illustrated in Table 13 show VP9 encoder to be significantly superior to the H.264/AVC encoder in sVQA ($p<0.025$) under low bitrates. VP9

encoder received ACR scores which are 80% higher than those of H.264/AVC encoder under 500kbps bitrate. However, the advantage of the VP9 encoder diminishes gradually while the bitrate increases. Under the highest bitrate of 6000kbps, the VP9 encoder only received 6.8% higher ACR scores than the H.264/AVC encoder.

Table 13. sVQA T-test results

Bit rate	Diff (%)	T-Test
500kbps	+80	t(11)=7.509, p<0.001
1000kbps	+27.8	t(11)=3.248, p=0.008
2000kbps	+14.3	t(11)=5.151, p<0.001
4000kbps	+8.9	t(11)=1.909, p=0.083
6000kbps	+6.8	t(11)=2.890, p=0.015
Average	+19.1	

The average subjective score collected from 30 participants for the highest bitrate (6000kbps) is about 7 points (9-point ACR scale is used). The subjective ACR scores revealed trends similar to the objective scores, either PSNR or SSIM. The H.264/AVC encoder appears to perform worse under 1080p full HD resolution than it does under 720p resolution.

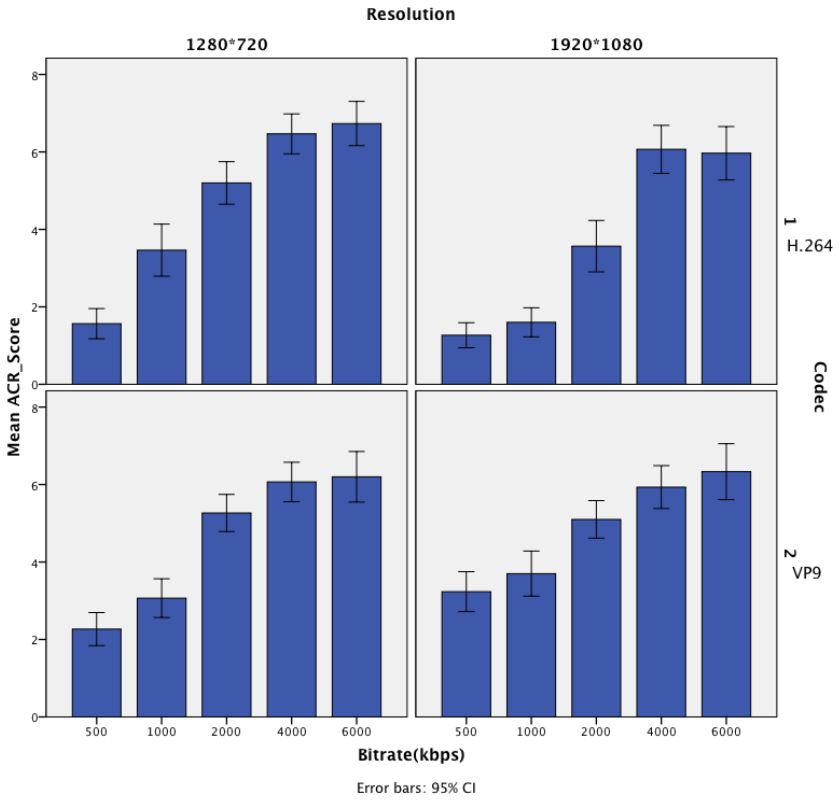


Figure 15. MOS of video content 1

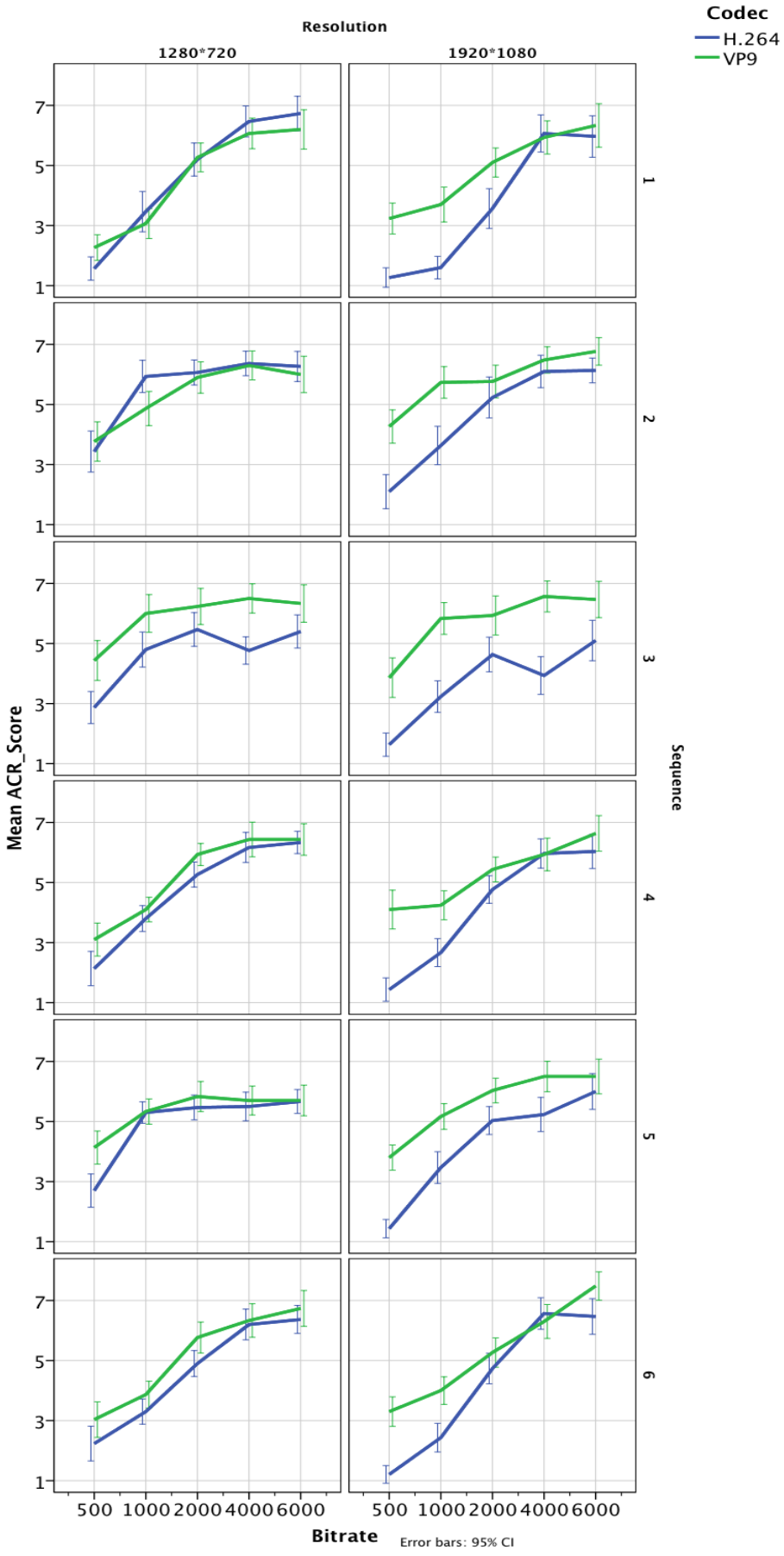


Figure 16. Average ACR scores separated by contents

Figure 16 shows clear trend: when the bitrate increases, the user perceived quality also increases. However, this correlation between bitrate increase and perceived quality improvement does not seem to be linear. The improvement of perceived video quality appears to saturate when the bitrate comes near to 6000kbps.

According to Figure 16, the difference in subjective assessment scores received for all six contents for the VP9 and H.264/AVC encoders under 720p resolution is significantly smaller than the difference shown for 1080p across all five bitrates. Under 720p resolution, the performances of both VP9 and H.264/AVC encoders are not as significantly different, as compared to 1080p resolution for all six types of contents, although VP9 is still clearly superior in both cases. However, for content 3, a low contrast scene with poorly lit background and foreground, the VP9 encoder performed significantly better than the H.264 encoder for both resolutions. Figure 12 also illustrates that VP9 encoded sequences received similar scores regardless of the spatial resolutions of the video sequences across all bitrates. On the contrary, the H.264/AVC encoder did not perform consistently for both resolutions. For all six types of contents, the ACR scores received for the H.264/AVC encoded video sequences under 1080p resolution are significantly lower than those under 720p resolution

4.3 OBJECTIVE ASSESSMENT

The outcomes of both PSNR and SSIM are highly consistent, showing similar patterns. Interestingly, the PSNR and SSIM scores shown in Figures 17 and 18 revealed that VP9 and H.265/AVC encoders have similar performances for both HD and full HD resolution. Video sequences encoded at different resolutions at the same bitrates received roughly the same PSNR and SSIM scores for both VP9 and H.265/AVC encoders. However, consistent with the sVQA scores, both the oVQA measurement metrics also show that the H.264/AVC encoder performed poorly under 1080p resolution, compared to its performance under 720p. As revealed by the PSNR scores averaged from all six video contents displayed in Figure 17, for achieving 32.5dB, the VP9 and H.265/AVC encoder need slightly more than 1000kbps bandwidth while the H.264/AVC encoder requires slightly more than 2000kbps of bandwidth.

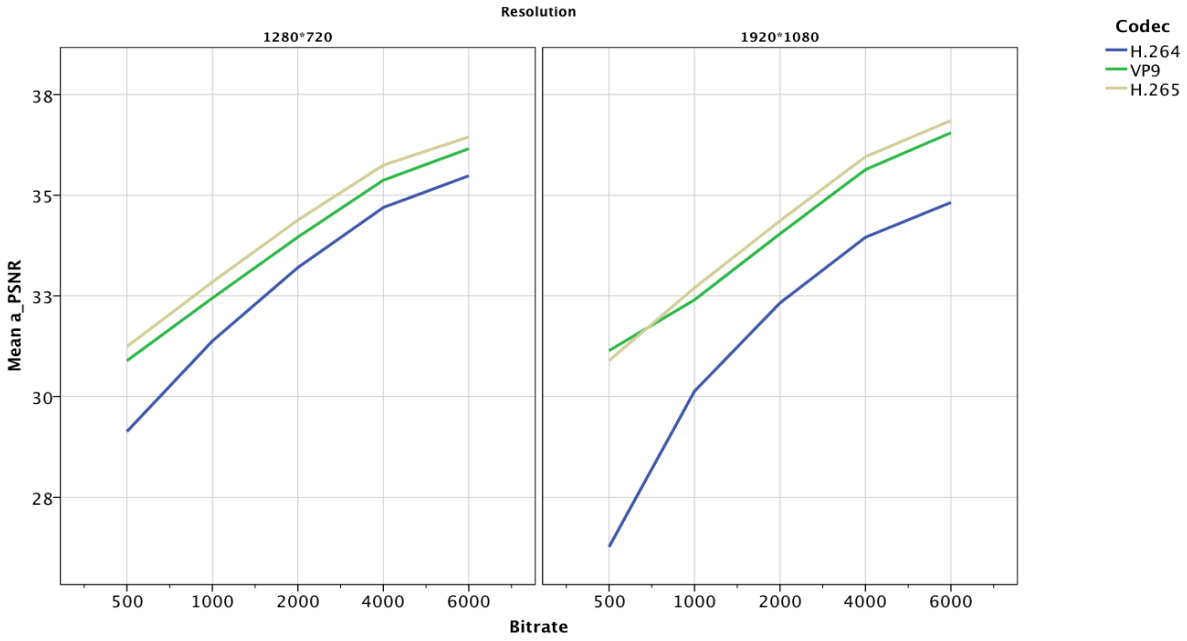


Figure 17. Averaged PSNR for 6 contents

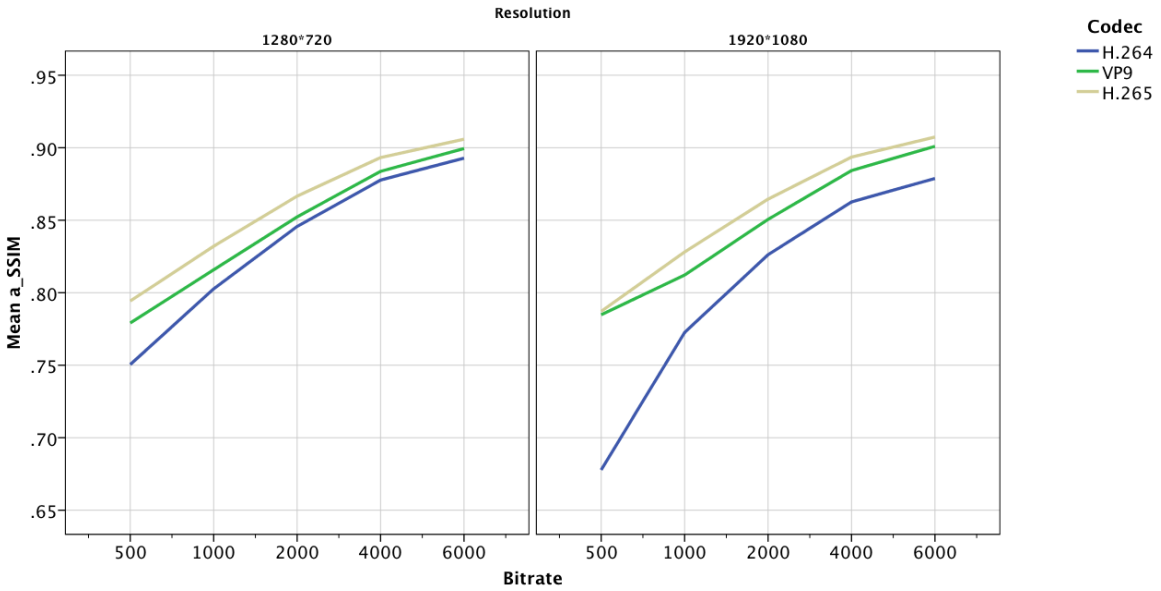


Figure 18. Averaged SSIM for 6 contents

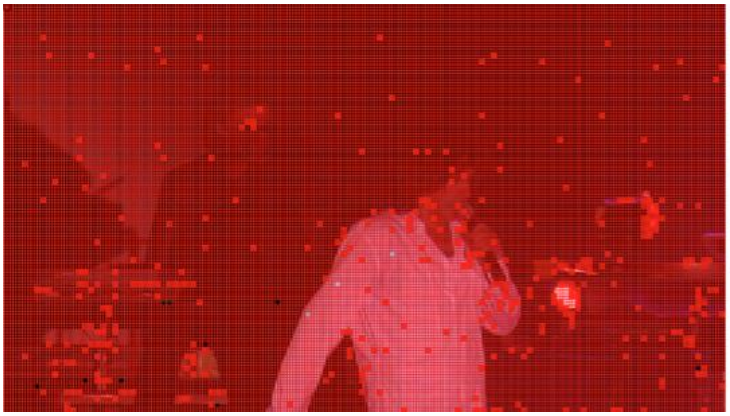
Figure 18 displays the averaged SSIM scores of all six video contents for three video encoders plotted against bitrates. These SSIM results show outcomes similar to the PSNR scores illustrated in Figure 17.

As we can observe from the PSNR and SSIM scores displayed in Figures 17 and 18, the performance of the H.264/AVC encoder suffers significantly under 720p resolution, compared to its performance under full HD 1080p resolution. This inconsistency in

performance for different image sizes corresponds to the Bjøntegaard metric scores and therefore is probably because the H.264/AVC encoder was designed in the early 2000s, before full HD video contents became popular. The designers of H.264/AVC encoders did not consider the spatial resolution factor in relation to the largest supported MB size and the extra MB overhead caused by the large number of MBs. The largest 16 × 16 MB specified in the H.264/AVC standard will have higher overhead under higher video resolution because the minimum theoretical number of MB will increase significantly once resolution increases. If more 16 × 16 MBs are required for a larger video image, video encoders will allocate more bits on defining the MBs, instead of storing actual video image information. Assuming that the size of video content in kilobytes remains constant, less actual video image information will be stored if too many MBs are defined by video encoders.

Bitrate in relation to the MB size also affects the performances of video encoders. Low bitrate video contents perform significantly worse than high bitrate ones because the ratio of MB overhead to the actual image information in lower bitrate video is higher than those higher bitrate videos. The larger the MB defined by video encoders, the less the MB overhead will be resulted because when the MB size gets bigger, less of it will be needed. Therefore, smaller MB size makes its performance inferior to itself in higher resolution as shown in Figures 19 and 20, generated by video sequence analysis software CodecVisa (CodecVisa, 2013). Based on the graphs, the H.264/AVC encoder underperformed significantly for 500kbps. When bitrate increases, it does not suffer as much comparing to itself although it is still significantly inferior to the 4th generation encoders because its largest supported size of MB is smaller.

If the H.264/AVC were to be used in any given situation, it is recommended that only HD 720p format only instead of using the full HD 1080p format.



MB Statistics

Type	Count	Bits	Bits/MB
I-MBs	8160	2168263	265.719
(0) I_NxN	8148	2165950	265.826
(14) I_16x16_1_0_1	3	317	105.667
(19) I_16x16_2_1_1	2	469	234.500
(20) I_16x16_3_1_1	1	302	302.000
(22) I_16x16_1_2_1	4	838	209.500
(24) I_16x16_3_2_1	2	387	193.500

Figure 19. H.264/AVC 6000kbps 1080p content MB Division

MB Statistics

Type	Count	Bits	Bits/MB
I-MBs	3600	1534074	426.132
(0)I_NxN	3600	1534074	426.132

Figure 20. H.264/AVC 6000kbps 720p content MB division

Figures 19 and 20 also show that significantly more MBs are needed by the H264/AVC encoder for 1080p resolution video sequences than for the 720p resolution. A total number of 8160 MBs are required for a given frame at 1080p resolution: 120 MBs per row multiplied by 68 MBs per column (120 × 68) equals 8160 MBs. Only 3600 MBs are needed for 720p resolution (80 × 46) equals 3600 MBs). It can be concluded 2.27 times more MBs are needed for full HD resolution than for HD resolution in a given video frame. The final file size of the video content is kept constant in this study. However, with the same given amount of bitrate, the full HD sequence requires significantly more MBs than HD to construct a video frame. It is speculated that the H.264/AVC encoder will spend more binary bits on the overhead of MB and therefore lower the amount of pixel / image information data in order to keep the bitrate constant. Therefore, H.264/AVC encoder performed significantly worse under 1080p resolution than 720p resolution in our oVQA.

Figures 21 and Figure 22 demonstrate the PSNR scores and SSIM scores of six video contents for all combinations of bitrates, resolutions and encoders. The SSIM and PSNR

scores are extremely content-type dependent. For content 2, encoded by H.265/AVC containing intensive motion and image details, the PSNR score ranges from about 35db to 40db across the bitrates from 500kbos to 6000kbps, whereas the range is about 25db to 33db for content 1, which contains extreme motion and detailed image. The less the motion and detail in the video sequences, the narrower the range. According to Figure 22, increasing the bitrate from 500kbps to 6000kbps will improve the SSIM score by only 0.05 for content 2 ; whereas the improvement can be as much as 0.16 for content 4.

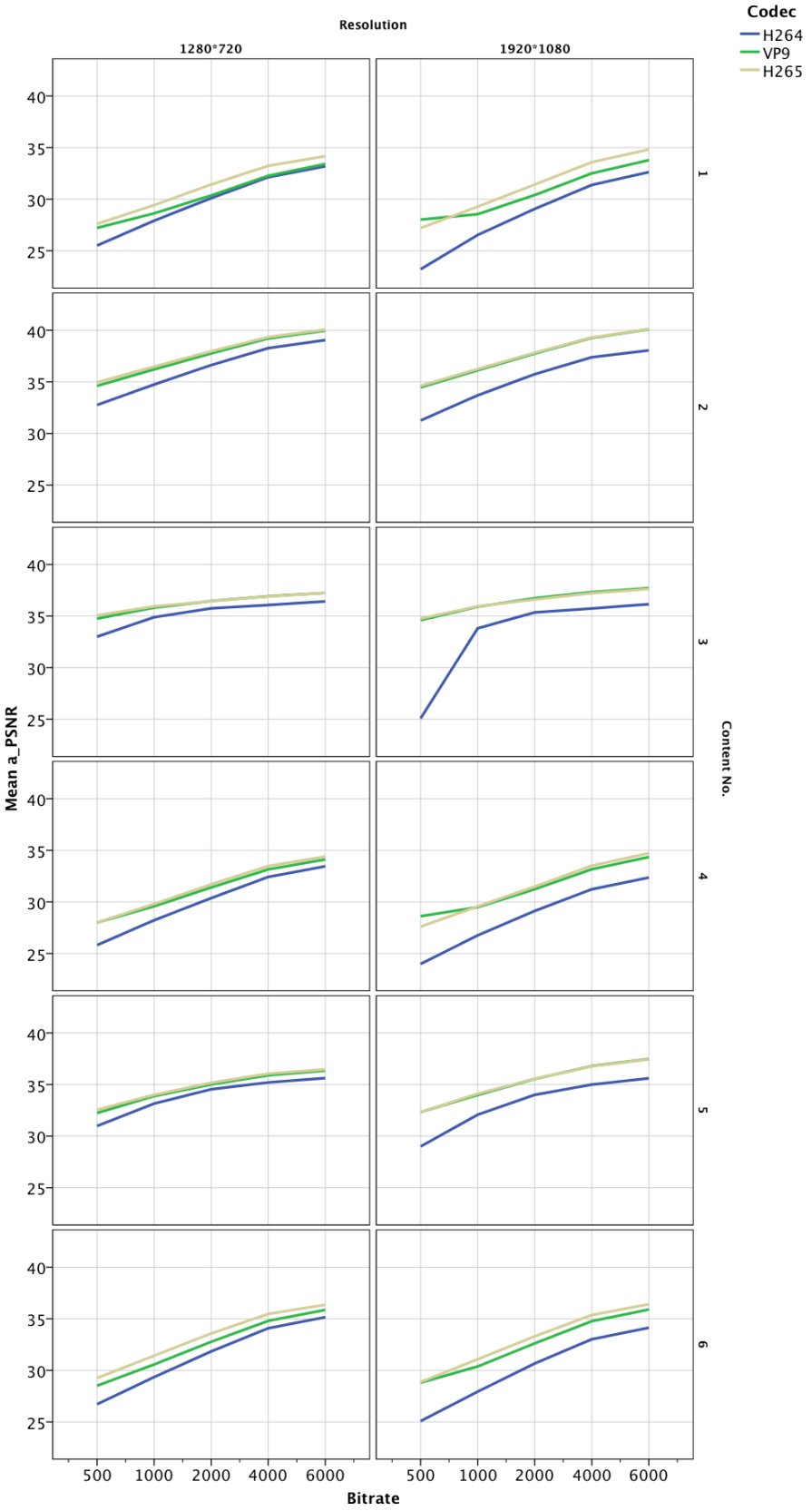


Figure 21. PSNR Scores of six contents for 720p and 1080p

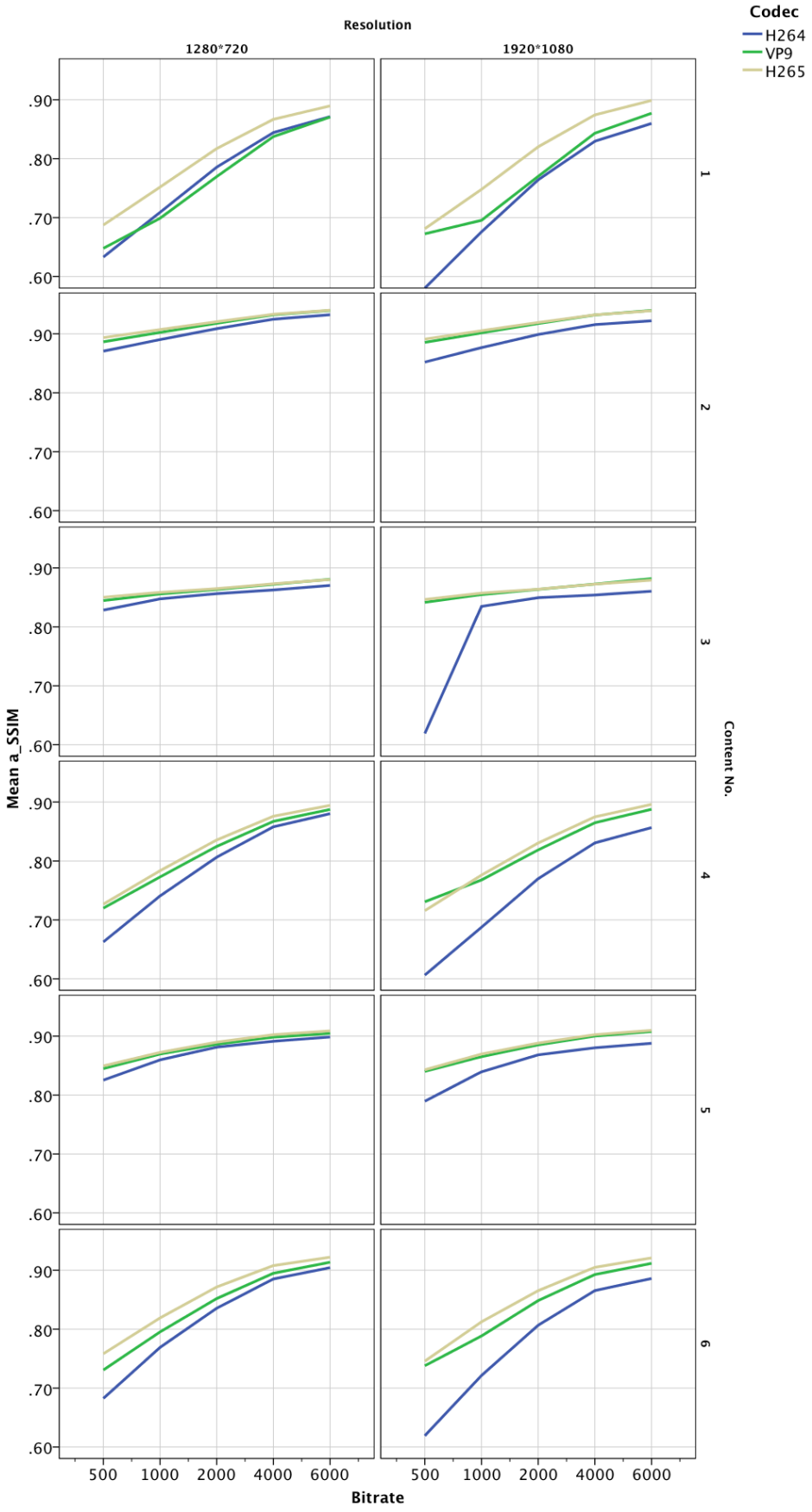


Figure 22. SSIM Scores of six contents for 720p and 1080p

It can be also observed from Figures 21 and 22 that both the VP9 and H.265/AVC encoders have almost identical performances achieving a visibly significant higher encoding efficiency than the H.264/AVC encoder across the lowest bitrate to the highest under both 1080p and 720p resolutions. Note that the H.265/AVC encoder performed significantly better than the VP9 and H.264/AVC encoders for video content 1 and video content 6, both of which consist of significant amounts of random moving objects. However, the advantage of the VP9 and H.265/AVC encoders over the H.264/AVC encoder is more obvious at a higher resolution (e.g., 1080p) and a lower bitrate (e.g., 500kbps). The overall improvement in objective scores of the VP9 and H.265/AVC encoders over the H.264/AVC encoder is from about 1.8% to 10% for PSNR and from 3.6% to 12% for SSIM. The advantage decreases as the bitrate increases. The significance of the superior performance of the VP9 and H.265/AVC over the H.264/AVC encoder, illustrated by the oVQA score, was confirmed by a set of paired t-tests for each pair of PSNR or SSIM scores between VP9, H.265 and H.264 under the same conditions (e.g., bitrate and resolution) ($p<.025$).

4.4 BITRATE SAVING

The Bjøntegaard metric is used to calculate the bitrate saving achieved by the test encoders, based on the PSNR scores. As five bitrates were implemented in this study, a five-point Bjøntegaard calculation was carried out. The results of the Bjøntegaard metric, shown in Table 14, revealed the average bitrate saving that the VP9 and H.265/AVC encoders achieved, relative to H.264/AVC encoder. Depending on the types of video contents, the H.265/AVC and VP9 encoders are up to 63.8% and 73.1% more effective than the H.264/AVC encoder for 720p and 1080p resolutions respectively. When spatial resolution increases, the performance edge of the H.265/AVC and VP9 encoders against H.264/AVC encoder will further increase.

Regardless of the spatial resolutions, both the 4th generation video encoders, H.265/AVC and VP9, outperformed H.264/AVC by a wide visible margin. Table 15 provides the bitrate saving of the H.265/AVC encoder over the VP9 encoder. The H.265/AVC encoder achieved up to about 26% of bitrate saving relative to VP9. The bitrate saving achieved by the H.265/AVC encoder is about 5% to 6% higher than the VP9 encoder under 720p resolution, and 3% higher under 1080p resolution. However, the H.265/AVC encoder achieved a significantly higher bitrate saving than that of VP9 for content 1 and content 6, which both contain extreme motions as well as image details for both spatial resolutions. For video

contents that do not contain fast random moving objects such as sequences 2, 3, 4 and 5, the efficiency advantage of the H.265/AVC encoder over the VP9 encoder is very marginal, ranging from 2% to 6% only. It is observed that for video content containing a dark background, such as sequence 3, the VP9 encoder outperformed the H.265/AVC encoder by 2.3% under full HD resolution. VP9 seems to have the performance advantage for high resolution video contents, with large amounts of black MBs.

Table 14. VP9 and H.265/AVC Bitrate saving compared to H.264/AVC

	VP9*	H.265/AVC*
720p25		
Video 1	13.94%	37.56%
Video 2	40.47%	46.16%
Video 3	57.04%	63.81%
Video 4	32.28%	38.36%
Video 5	31.84%	37.57%
Video 6	27.07%	43.12%
1080p25		
Video 1	35.46%	52.83%
Video 2	57.83%	59.79%
Video 3	70.38%	73.13%
Video 4	53.49%	56.38%
Video 5	54.45%	56.03%
Video 6	46.25%	55.88%

* Rounded off to 2 decimal places

Table 15. H.265/AVC Bitrate saving compared to VP9

	H.265/AVC*
720p25	
Video 1	28.91%
Video 2	9.74%
Video 3	0.33%
Video 4	9.33%
Video 5	9.23%
Video 6	23.08%
1080p25	
Video 1	25.85%
Video 2	3.93%
Video 3	-2.32%
Video 4	5.90%
Video 5	1.79%
Video 6	18.06%

*Rounded off to 2 decimal places

The VP9 and H.265/AVC are both considered to be the latest generation of encoders. However, based on the bitrate saving calculation, the H.265/AVC encoder is slightly more efficient than the VP9 encoder, especially for encoding contents consisting fast-moving objects.

4.5 BITRATE SAVING RELATIVE TO MICRO-BLOCK

According to the Bjøntegaard metric scores tabulated in section 4.4, both the VP9 and the H.265/AVC encoders are significantly superior to the H.264/AVC encoders. However, the VP9 and H.265/AVC encoders are capable of achieving a 10% to 12% higher bitrate saving under 1080p resolution comparing to their saving under 720p resolution. This obvious inconsistency in performance is speculated to be caused by the larger MB sizes supported by the 4th generation video encoders. The largest MBs defined in the VP9 and H.265/AVC encoders are 64 × 64 and 16 × 16 in H.264/AVC respectively. When the video frame gets larger in term of spatial resolution, VP9 and H.265/AVC encoders will define large 64 × 64 MBs and therefore do not need to define a higher number of small MBs that cause encoding overhead, and thus superior bitrate saving is achieved. Different from the 4th generation video encoders, much smaller MBs of 16 by 16 pixels are compulsory to be defined by H.264/AVC encoders even in the situations that image complexity and motion level are low for a given video content.

Additionally, a significantly larger theoretical minimum number of MBs will be defined in higher spatial resolution video content for the H.264/AVC encoder than for 4th generation encoders. The H.264/AVC encoder has to define theoretically 8160 MBs minimum for the 1080p video frame and 3600 MBs minimum for the 720p video frame. On the contrary, 4th generation video encoders such as VP9 and H.265/AVC have to define theoretically 510 MBs minimum for 1080p video frame and 240 MBs minimum for 720p video frame. In a real-life situation, it is not possible for all three video encoders to define the theoretical minimum number of MBs; the number of MBs in a given video frame would be much higher than the theoretical minimum. However, the 4th generation video encoders have the capability to define far fewer numbers of MBs in a video frame than the H.264/AVC encoder depending, on the complexity of video image.

Comparing the 4th generation video encoders head to head is worthwhile as video industries have to make decisions on which to adopt in the near future. Both the VP9 and H.265/AVC encoders have minimal design differences, they performed slightly differently

according to the bitrate saving results. H.265/AVC is capable of achieving slightly higher bitrate saving, compared to VP9: approximately 2% to 5% in 1080p resolution and 10% in 720p resolution. According to Table 2 (see Chapter 2), the VP9 encoder is capable of breaking down 64 × 64 SB into smaller rectangular sub-blocks for more effective allocating of MBs. It is speculated that either the H.265/AVC encoder outperforms or the VP9 encoder underperforms under higher resolution. Further investigation is required to explain this observation; however, this is beyond the scope of this study.

The H.265/AVC encoder performed significantly better than the VP9 by achieving approximately 25% of bitrate saving for video contents 1 and 6 which contain fast random motion and detailed foregrounds and backgrounds. Content 1 contains large groups of fast-moving marathon runners, while content 6 contains randomly moving actors in the foreground in a gradually panning and zooming scene. Such significant superiority in quality is most likely achieved because of the higher number of motion vectors supported in the H.265/AVC encoder (Sullivan et al., 2012). The H.265/AVC encoder supports 33 motion vectors, compared with 8 by the VP9 encoder.

4.6 CORRELATION

PCC was conducted for all three video encoders. The outcome of subjective assessment is highly consistent and correlates with objective study outcome. The PCC data presented in Tables 16 to 18 illustrate the PCC between the PSNR, SSIM and ACR scores achieved by all three video encoders. The correlation between the scores is extremely strong and linear. However, although there is significant correlation between subjective scores and objective scores (at < 0.01 level 2-tailed), the correlation does not seem to be linear as PCC value ranges from 0.5 to 0.6.

Table 16. H.264 sVQA and oVQA PCC

		PSNR_Score	SSIM_Score	ACR_Score
PSNR_Score	Pearson Correlation	1	.950**	.600**
	Sig. (2-tailed)		.000	.000
	N	1800	1800	1797
SSIM_Score	Pearson Correlation	.950**	1	.638**
	Sig. (2-tailed)	.000		.000
	N	1800	1800	1797
ACR_Score	Pearson Correlation	.600**	.638**	1
	Sig. (2-tailed)	.000	.000	
	N	1797	1797	1797

**. Correlation is significant at the 0.01 level (2-tailed).

Table 17. VP9 sVQA and oVQA PCC

		PSNR_Score	SSIM_Score	ACR_Score
PSNR_Score	Pearson Correlation	1	.920**	.494**
	Sig. (2-tailed)		.000	.000
	N	1800	1800	1796
SSIM_Score	Pearson Correlation	.920**	1	.523**
	Sig. (2-tailed)	.000		.000
	N	1800	1800	1796
ACR_Score	Pearson Correlation	.494**	.523**	1
	Sig. (2-tailed)	.000	.000	
	N	1796	1796	1796

Table 18. H.265/AVC sVQA and oVQA PCC

		PSNR	SSIM
PSNR	Pearson Correlation	1	.921**
	Sig. (2-tailed)		.000
	N	60	60
SSIM	Pearson Correlation	.921**	1
	Sig. (2-tailed)	.000	
	N	60	60

**. Correlation is significant at the 0.01 level (2-tailed).

According to Tables 16, 17 and 18, SSIM achieves a slightly higher correlation with the ACR than that of PSNR. The curve estimation of PSNR, SSIM and ACR shown in Figures 23, 24 and 25 confirmed that the correlation between ACR and SSIM or PSNR is non-linear and appears to be logistic or logarithmic; the correlation between SSIM and PSNR is highly linear.

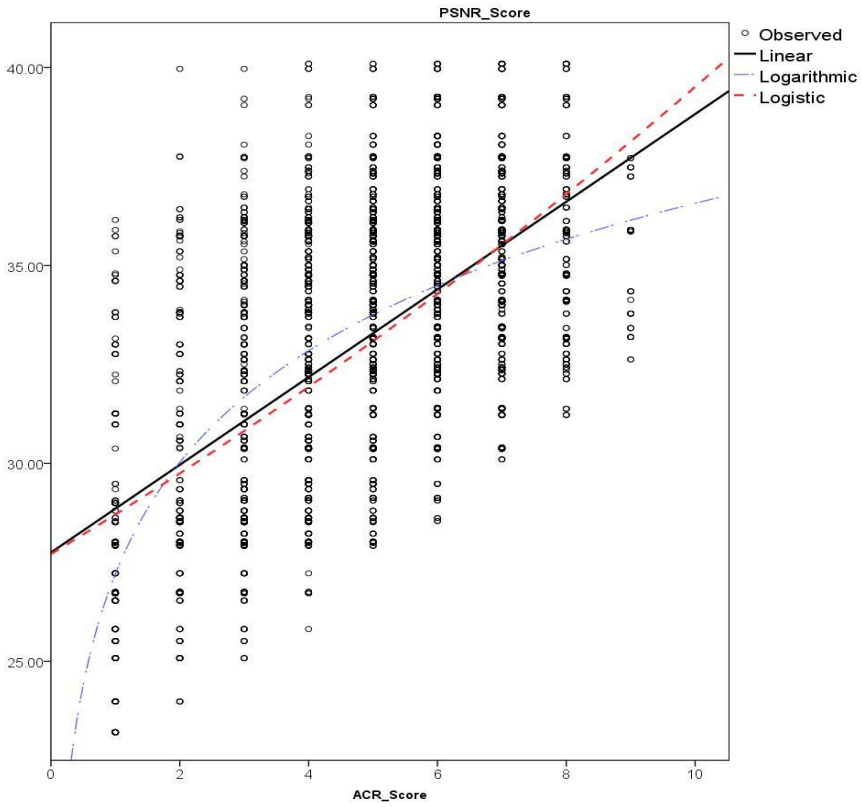


Figure 23. ACR and PSNR curve estimation

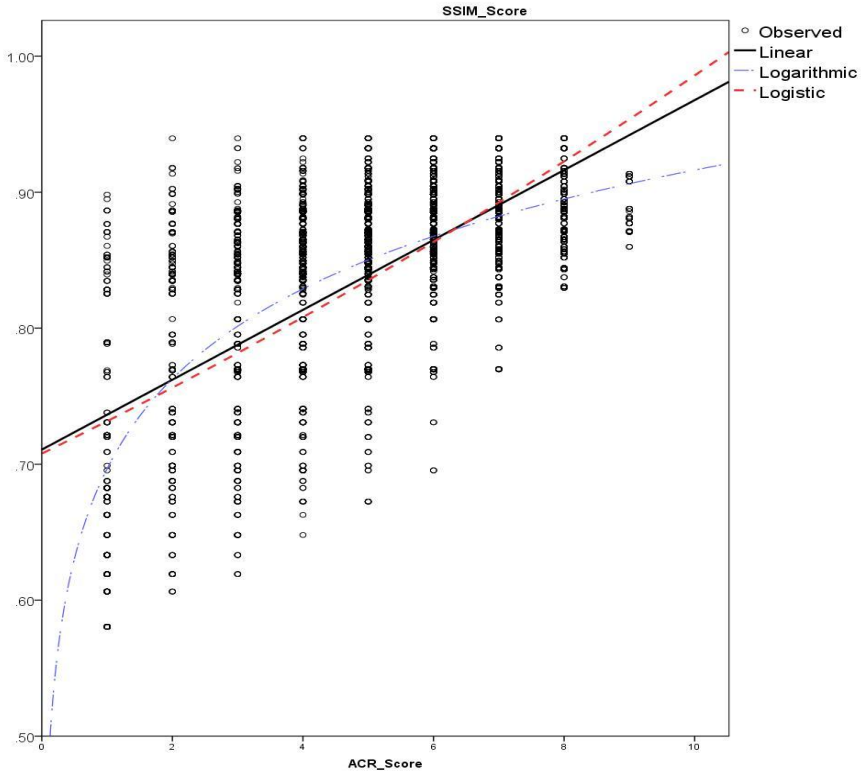


Figure 24. ACR and SSIM curve estimation

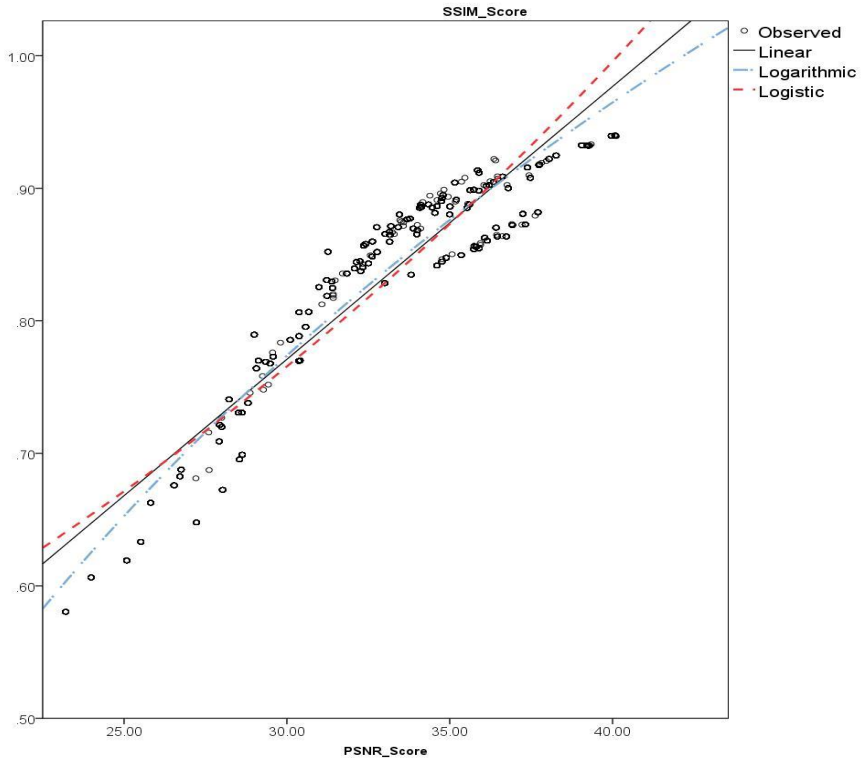


Figure 25. PSNR and SSIM curve estimation

4.7 DISCUSSION

There is a potential impact on the subjective assessments caused by the order of showing the distorted video sequences to the participants, as a previous study suggested participants' ratings of a particular video sequence depend on the perceived quality of the previous video sequence watched, the so called memory effect (Matthews & Stewart, 2009). In this study, the nature of the ACR method means it is not possible for us to completely eliminate the memory effect or to quantify the influence caused by it. However, we randomized the order of video sequences for each session in an effort to reduce this effect; it is recommended to randomize the displaying order of distorted sequences for each participant to eliminate the memory effect for any future studies.

Additionally, the average highest ACR score obtained from six video contents is about 7-point even for the videos under the highest bitrate of 6000kbps. It is possible the bitrate of 6000kbps shown on 10 inch small factor full HD screen is relative low to the screen size. Mobile devices with different screen sizes should be tested in the future studies.

It is observed in this study that, when bitrate increases, the perceived video quality and objective quality increase correspondingly in a non-linear monotonic logistic or logarithmic

fashion. In order to get a linear correlation between bitrate and sVQA and oVQA scores, the bitrate can possibly be made logarithmic or logistic as a predictor in the prediction models.

We have also successfully displayed that the type of video content will significantly affects both subjective and objective assessment outcomes. Video contents containing extreme motions and image details will receive lower subjective and objective scores than video contents with minimum motion and relatively simple image details under the same bit rate. The definition of video content plays an important part in estimating the perceived quality of end users. Therefore, predictors which define video contents should be primarily focused when creating prediction models.

Chapter 5: Perceived Video Quality Modelling

This chapter contains detailed information about predictors and their selection, categories, and testing, towards proposing user-perceived video quality prediction models on small-form factor screens.

5.1 PROPOSED PREDICTORS

The proposed predictors aim to define both the content type and the nature of the encoders. Three categories of predictors are considered in this study: the definition of video content, the definition of video encoder and the encoding parameter.

Table 19. Proposed predictors

Category	Parameters
Content Definition	Resolution, total Pixel, Motion level, complexity
Encoding Parameter	Bitrate, scanning mode, frame rate, *Chroma* and *Luma* sampling, frame structure
Encoder Definition	MB Size, Minimum MB Number Per Frame, Intra and Inter MB ratio

Table 19 illustrates the three proposed predictor categories. Factors such as scanning mode and frame rate are set as constant to align with the scopes of this study, so only the bitrate of the encoding parameter category, the predictors of content and the encoder definition categories are tested for their prediction accuracy using IBM SPSS.

Figure 26 displays the PSNR and ACR scores for all six video sequences across five bitrates. Both subjective and objective studies revealed that the type of video content plays an important role in the scores received. For instance, video content 2, which consists of only close-up portrait scene with minimum motion and image details, therefore received much higher ACR and PSNR scores for all five bitrates, than scored by video content 1, which included fast motion and a complicated video scene. It can thus be concluded that video content definition, the key category of predictors, should be considered for prediction model.

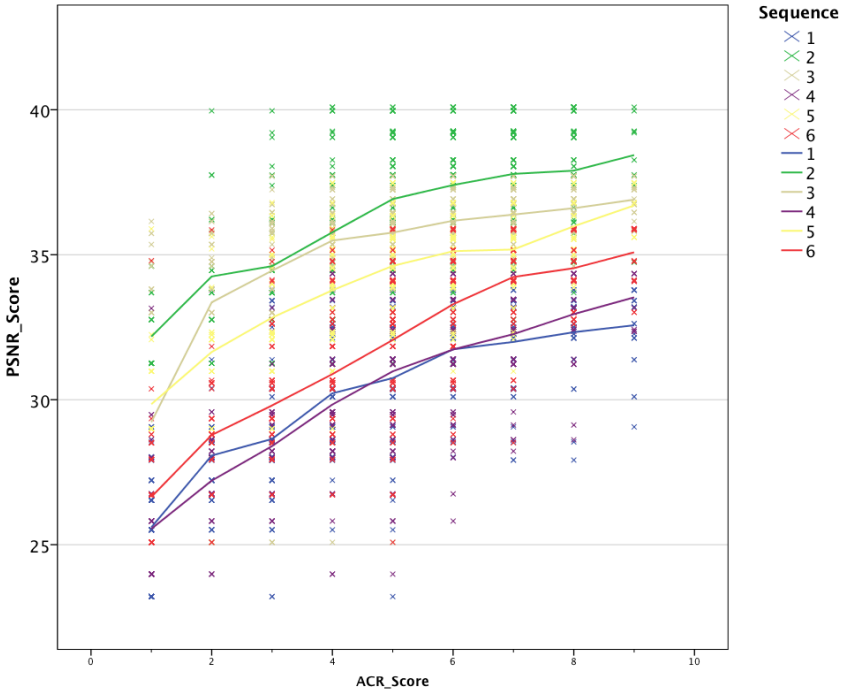


Figure 26. PSNR and ACR scores for different contents

The definition of video encoders should be considered, as the second most important category of predictors. In this study, as well as making use of the efforts of other researchers to define video contents in various ways, we also propose the idea of defining types of video encoders as well. Whenever there is a new video encoder, new characteristic values can be plugged in. The fundamental mechanism and characteristics are the keys. This not only provides a solution, taking the type of video encoders as a category of predictors in the creating of the prediction model, but also makes the prediction model future proof to a certain extent. For instance, if a new generation of video encoders is created, researcher can extend the application of the prediction model created in this study to evaluate the performance of any new video encoders by plugging in the new values which define them.

Content Definition: Since both 720p and 1080p video sequences were used in our study, it would be necessary to firstly define the spatial resolution. Due to the fact that modern digital video sequences are represented in binaries, the higher the spatial resolution, the higher the amount of binaries needed to represent a video frame. Video spatial resolution is also represented in the number of horizontal pixels multiplies by the number of vertical pixels. Therefore, the spatial resolution predictor adopted in this study is the total pixel count in a given video frame. For instance, 720p video sequences will have a spatial resolution represented, in total pixel number in a given video frame, by 1280 × 720, which equals

921600 pixels, while 1080p video sequences will have 1920 × 1080, which equals 2073600 pixels. The predictor values were normalized to a ratio of 2.25. Besides using the total pixel number ratio to define the spatial resolution of test video sequences, we also considered the spatial and temporal complexity of video frames to define the types of video contents. The average Spatial Information (SI) and the average Temporal Information (TI) over all frames of a video were calculated using the method defined in ITU document (ITUR Rec, 2012a).

Encoding Parameter: First of all, predictors which were set as constant or beyond the control of researchers were excluded from this study. For example, *Chroma* and *Luma* sampling accuracies were standardized as YUV4:2:0, the encoding passes were set to 1-pass only and the frame rates were fixed at 25 frames per second in this study, and therefore these constants were not taken into consideration as predictors. On the contrary, the bitrate were taken as a predictor, with its values normalized to a single digit by being divided by one thousand. For instance, a bitrate of 500kbps will be 0.5. However, since it was observed there is a logarithmic or logistic correlation between the sVQA or oVQA scores and the five different bitrates, according to the data illustrated in Chapter 3, we introduced the logarithmic bitrate values, along with the original values after normalization to create a linear model.

It is worth noting that encoders such VP9 were not designed to have a fixed GOP structure as H.264/AVC and H.265/AVC encoders. It is technically impossible to use values such as GOP length or related GOP structure information as predictors in this study.

Encoder Definition: According to the literature review, the main differences setting the latest generation video encoders apart from the previous ones are the support maximum sizes of MB. Both the VP9 and H.265/AVC encoders support MB as large as 64 × 64 pixels, while the H.264/AVC encoder supports only the largest MB of 16×16 pixels. We took this into consideration as the predictors that can define the characteristics of video encoders. This is a future proof approach, as new video encoders can always be defined by using their characteristics, and therefore a uniform prediction model can always be used for better speculated accuracy and consistency. We begin by taking the total pixel number ratios of the supported largest MB and normalized them: 4096 pixels versus 256 pixels were normalized to 16 and 1 for the H.265/AVC / VP9 encoder and the H.264/AVC encoder respectively. Total number of supported largest MB for different spatial resolutions were calculated. For example,

the theoretical minimum number of MB needed for 1080p resolution video sequences encoded by H.264/AVC is 8160 while it will be only 240 for H.265/AVC and VP9 encoders for 720p resolution video sequences. The calculation of the theoretical minimum number of MB can be expressed as such: (number of horizontal pixels ÷ supported largest MB pixel length) × (number of vertical pixels ÷ supported largest MB pixel length). For normalization, the number is then divided by 100.

Although a ratio of intra and inter MB number or bytes in a given video sequence can be a potentially strong and logical proposition as predictor, technical constrains mean that the related statistics cannot be generated by various bit stream analyzers available, because of the capability issues caused by video sequences encoded by different versions of H.265/AVC HM test model encoders.

It has been also attempted to use the combination of predictors as new predictors in this study: for instance, the division or multiplication of ATI and ASI. However they are proven to be insignificant in our data analysis software.

5.2 PROPOSED VIDEO QUALITY PREDICTION MODEL

The ACR scores by 30 participants were set as the dependent in IBM SPSS and all the predictors were loaded as the independents. Tables 20 and 21 illustrate the stepwise predictor test by linear regression in SPSS. It is observed that available predictors can achieve the highest accuracy of 91.5%. It should be noted that no predictors derived from oVQA were used in the model creation because PSNR and SSIM produce scores that are relative values which are generated from the comparison between the original video sequences and the distorted ones. In many real life situations, the original (uncompressed) video sequences are usually not available to be used as reference video.

For the modeling, the variables used include:

- BR stands for the true bitrate in kbps that is divided by 1000
- LBR stands for the logarithmic bitrate with base 10
- M stands for the ratio of the total number of the pixels in a video frame of given spatial resolution to the maximum number of MB can be possibly allowed, and then normalized by 100.
- Pixel_Ratio represents the ratio of total number of pixels in a video frame divided by 1280 × 720 pixels (720p resolution). E.g., Pixel_Ratio equals to 1 in a 720p video while it equals to 2.25 in a 1080p video.
- TI denotes the average value of temporal information over all frame of a video.

- SI denotes the average value of Spatial Information over all frames of a video.

In our model test without objective scores available, both SI and Pixel_Ratio predictors were excluded by the step-wise regression model test carried out in SPSS.

Table 20. Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.825[a]	.680	.678	.84365832
2	.890[b]	.792	.789	.68318322
3	.904[c]	.817	.812	.64412884
4	.915[d]	.837	.831	.61016555

a. Predictors: (Constant), LBR

b. Predictors: (Constant), LBR, M

c. Predictors: (Constant), LBR, M, BR

d. Predictors: (Constant), LBR, M, BR, TI

Table 21. Model coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-5.392	.648		-8.318	.000
	LBR	3.113	.196	.825	15.845	.000
2	(Constant)	-4.924	.528		-9.321	.000
	LBR	3.113	.159	.825	19.567	.000
	M	-.042	.005	-.334	-7.934	.000
3	(Constant)	-9.911	1.357		-7.306	.000
	LBR	4.940	.486	1.309	10.166	.000
	M	-.042	.005	-.334	-8.415	.000
	BR	-.369	.093	-.509	-3.952	.000
4	(Constant)	-9.270	1.296		-7.152	.000
	LBR	4.940	.460	1.309	10.732	.000
	M	-.042	.005	-.334	-8.883	.000
	BR	-.369	.088	-.509	-4.172	.000
	TI	-.028	.008	-.142	-3.778	.000

a. Dependent Variable: Avg_ACR

As illustrated by Table 21, 4 predictors are needed for model 4 to achieve the highest possible prediction accuracy of 91.5% while lower accuracy can be achieved by using less predictors. The four predictors are BR, LBR, M and TI respectively. The predictor of the SI

has minimal significance in the models. However, the theoretical minimum number of MB required in a given spatial resolution is significantly relevant as shown in the model test. The theoretical minimum number of MB required in given spatial resolution correlates spatial resolution of video frame and the largest MB size supported by video encoder. If the prediction accuracy were to be compromised for less predictors are available, prediction accuracy of 90.4%, 89% and 82.5% can be achieved by using 3, 2 and 1 predictor respectively. Logarithmic bitrate is the most significant predictor, followed by the theoretical minimum number of MB required in a video frame.

Prediction accuracy decreases sharply from 89% to 82.5% if only 1 predictor instead of 2 were to be used as shown by the suggested prediction model 1 and 2. Comparing to the suggested model 4, suggested model 3 does not take temporal information of video sequences as predictor and therefore achieved prediction accuracy that is 1.1% lower. Based on suggested model 4 of the model test shown by Table 20 and 21, the proposed user perceived video quality on small-form factor screen without objective scores is as (1):

$$S_{ACR} = LBR \times 4.94 - M \times 0.042 - BR \times 0.369 - TI \times 0.028 - 9.27 \quad (1)$$

LBR is the logarithmic bitrate; M represents the theoretical minimum number of CTU or MB needed in a video frame that correlates with spatial resolution of video frame and maximum size of MB supported by video encoders; BR represents the bitrate and the value 9.27 is constant.

The R value for model (1) if 0.915, indicating a close correlation between the predicted ACR scores and the scores obtained by sVQA. Adjusted R^2 value that measures the goodness-of-fit for a model is 0.831, indicating 83.1% of the variations in ACR is captured by model (1). The standard error is relatively low at about 0.6. Figure 27 shows the scatter-plot of the predict values versus the average sVQA scores.

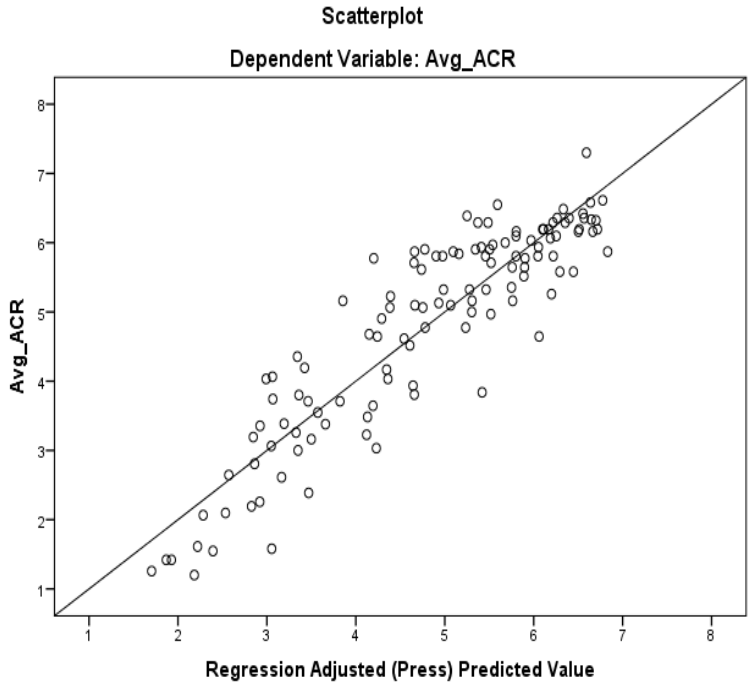


Figure 27 Predicted ACR without objective scores

The proposed model (1) does not take oVQA scores such as SSIM and PSNR into consideration. In the situation that reliable oVQA scores such as PSNR and SSIM are available, predictors derived from these scores can also be used to improve prediction accuracy. Assuming both PSNR and SSIM scores are available, Tables 22 and 23 show the improved prediction accuracy.

Table 22. Model summary with SSIM scores

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.825[a]	.680	.678	.84365832
2	.905[b]	.819	.816	.63798973
3	.936[c]	.876	.873	.52908213
4	.945[d]	.893	.889	.49451100

a. Predictors: (Constant), LBR

b. Predictors: (Constant), LBR, SSIM

c. Predictors: (Constant), LBR, SSIM, M

d. Predictors: (Constant), LBR, SSIM, M, BR

Table 23. Model coefficients with SSIM scores available

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-5.392	.648		-8.318	.000
	LBR	3.113	.196	.825	15.845	.000
2	(Constant)	-8.899	.615		-14.475	.000
	LBR	1.926	.195	.510	9.896	.000
	SSIM	8.885	.940	.487	9.452	.000
3	(Constant)	-7.891	.528		-14.948	.000
	LBR	2.148	.164	.569	13.084	.000
	SSIM	7.219	.812	.396	8.893	.000
	M	-.031	.004	-.250	-7.357	.000
4	(Constant)	-11.865	1.064		-11.155	.000
	LBR	3.695	.398	.979	9.294	.000
	SSIM	6.898	.763	.378	9.045	.000
	M	-.032	.004	-.254	-7.986	.000
	BR	-.304	.072	-.419	-4.217	.000

a. Dependent Variable: Avg_ACR

PSNR predictor was excluded by the step-wise regression model building in SPSS. SSIM scores were selected instead of PSNR scores is probably because SSIM scores range from 0 to 1.0 with 0 being the lowest possible score and 1 represents the highest possible score. On the contrary, the decibel values of PSNR scores are not so uniformed and do not have a fixed range. Therefore the prediction model of perceived video quality on small-form factor screen proposed with the presence of objective assessment scores is as follows:

$$S_{ACR} = LBR \times 3.695 - M \times 0.032 - BR \times 0.304 + SSIM \times 6.898 - 11.865 \quad (2)$$

With SSIM score available, TI is no longer needed as a predictor. The highest prediction accuracy achieved by the model test is 94.5% with four predictors which are SSIM, LBR, BR and M. If less predictors were to be used, prediction accuracy will decrease to 93.6%, 90.5% and 82.5% with 3, 2 and 1 predictors used respectively. LBR is the most significant predictors among the 4 and the rest are SSIM, M and BR respectively.

The R value for model (2) if 0.935, indicating a close correlation between the predicted ACR scores and the scores obtained by sVQA. Adjusted R^2 value that measures the goodness-of-fit for a model is 0.889, indicating 83.1% of the variations in ACR is captured by model (2). The standard error is relatively low at about 0.5. Figure 28 shows the scatter-plot of the predict values versus the average sVQA scores.

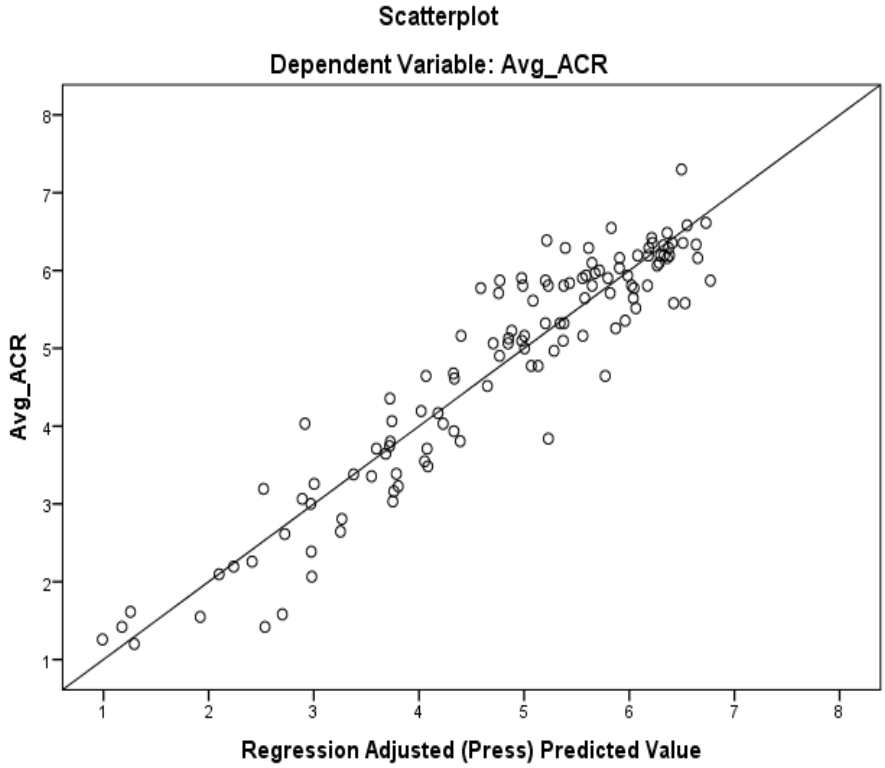


Figure 28 Predicted ACR with SSIM scores

The two prediction models proposed achieved the prediction accuracy of 91.5% to 94.5% by using four predictors each (bitrate, logarithmic bitrate, minimum theoretical number of MB in a frame and TI or SSIM score). The prediction accuracies achieved for both model (1) and (2) are the best possible accuracies even in the situations where more predictors were to be used.

5.3 SUBJECTIVE SCORE PREDICTION FOR H.265

The same set of predictors which define video sequences encoded by the H.265/AVC and VP9 encoders have exactly the same values, due to their similarity in design. Therefore, the proposed model (1) and model (2) will predict exactly the same ACR scores for video sequences encoded by both H.265/AVC and VP9 encoders. However, both sVQA and oVQA outcome of this study revealed that the H.265/AVC encoder is marginally more efficient than

the VP9 encoder. The prediction models proposed can be enhanced by differentiating the two encoders by their oVQA scores.

In order to enhance the prediction accuracy of the proposed sVQA prediction model that takes the oVQA score into consideration, for the H.265 encoded video specifically, we propose a sub-model to predict the SSIM value of video encoded by the H.265/AVC encoder, based on the SSIM value of the same video encoded by the VP9 encoder. Although we have generated the oVQA scores for H.265 encoder in this study, this sub-model allows H.265 SSIM scores to be predicted based on the available SSIM scores without spending any time and resources to actually encode video sequences to the H.265 format. The sub-model achieved 98.3% of accuracy in predicting the SSIM score of H.265 encoded videos by regression as follows:

$$\boldsymbol{SSIM_{H265} = SSIM_{VP9} \times 0.885 + 0.108} \quad \textbf{(3)}$$

Table 24 Model summary of predicting H.265 SSIM

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.983[a]	.966	.966	.01173986

a. Predictors: (Constant), Bitrate, SSIM_VP9

Table 25 Model coefficients of predicting H.265 SSIM

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.108	.018		5.870	.000
	SSIM_VP9	.885	.022	.983	40.676	.000

a. Dependent Variable: SSIM_H265

Hence, the ACR scores of the video sequences encoded by the H.265/AVC encoder can be calculated by substituting the SSIM values achieved by model (3) into model (2). For example:

$$\boldsymbol{S_{ACR\,H265} = LBR \times 3.695 - M \times 0.032 - BR \times 0.304}$$
$$\boldsymbol{+ (SSIM_{VP9} \times 0.885 + 0.108) \times 6.898 - 11.865}$$

The SSIM value of video encoded by the H.264/AVC encoder can also be used to predict the SSIM scores of the same video encoded by the H.265/AVC encoder. However, H.265 as the latest generation of video encoder shares more design similarities with the VP9 encoder. Therefore we used SSIM values generated by the VP9 encoder instead of the H.264/AVC encoder to predict the SSIM value generated by the H.265/AVC encoder.

5.4 SUMMARY

Two prediction models are proposed based on this study. Model (2) that makes use of the objective score achieved the prediction accuracy of 94.5% by using four predictors which are bitrate, logarithmic bitrate, ratio of frame pixel and size of support MB, along with the SSIM score respectively. Model (1) does not take the objective score and is capable of achieving a prediction accuracy of 91.5% by replacing the SSIM score by the average temporal information of the video. Spatial resolution and the largest MB size supported by the video encoder do not significantly affect the end user perceived video quality. Instead, the correlation between them does. Besides, the bitrate and the logarithmic bitrate derived from the encoding parameter category of predictors and the average TI of content definition category of predictors also significantly affect the perceived video quality of end user on small-form factor screen. The two models can be used according to situations when the objective measurement is available.

Additionally, a sub-model is also created to further enhance the prediction accuracy of the proposed model (2) by differentiating the SSIM scores of the VP9 and H.265/AVC encoded videos. Furthermore, sub-model indirectly allows prediction of the subject score (ACR) of H.265 video sequences without having to encode to H.265 format.

Chapter 6: Conclusions

This chapter outlines this study's contribution, its associated limitations, and its recommendations for related future works. The research questions and goals proposed in Chapter 1 are answered, as illustrated by the prediction accuracy of the proposed models.

6.1 SUMMARY OF KEY FINDINGS

The research goals outlined in Chapter 1 are fulfilled. In this study, the actual end user perceived video quality on mobile device was discovered by the subjective approach of ACR; SSIM and PSNR are the objective approaches. The outcomes from all approaches are consistent, as shown by the PCC scores, and 91.5% to 94.5% of prediction accuracy is achieved. Depending on the type of video content, the latest generation of video encoders, VP9 and H.265, are about 50% more efficient than the H.264/AVC encoder on mobile devices. Such a significant amount of performance improvement is mainly attributed by the designed structures of video encoders. The H.265 and VP9 encoders allow a much larger size of MB, compared to H.264/AVC encoder. Therefore, the supported largest size of MB becomes a potent predictor in the proposed performance prediction models. Other predictors include logarithmic bitrate, bitrate, and oVQA score.

By plugging the available predictors, which can be obtained easily, the end user perceived video quality for any given video sequence encoded by the H.265/AVC and VP9 encoders can be calculated from the proposed models. The oVQA sub-model proposed also allows estimating of the oVQA score (SSIM predictor) to calculate the perceived video quality without actually having to spend time and resources encoding video sequences with the H.265/AVC encoder.

6.2 CONTRIBUTION

The proposed models can be embedded into the encoding and distributing strategies of video content distributors to manipulate the perceived video quality and service on small-form factor screens. Mobile devices are becoming very popular, so more and more users will consume video content on mobile devices instead of on conventional TV system. Content distributors will need the proposed prediction models to adapt to the new business environment caused by this dramatic change in user behavior. With the proposed prediction

models in this study, video content distributors will not only be able to improve QoE while maintaining their operational expenses, but will also maintain the current level of QoE while reducing their operation cost.

The proposed prediction models provide a new aspect on how video quality prediction models can be built with researchers by taking the correlation between spatial resolutions and the largest supported MB of video encoders as a predictor. This makes the models encoder-independent in the prediction models that predict perceived video quality. As long as the largest supported MB of any video encoder is known, the models can use this information to define video encoders and hence predict the perceived video quality on small-form factor screens with high accuracy. This is unprecedented in how video quality prediction models are built as previous researches do not attempt to define video encoders according to their characteristics. To researchers, the proposed models can also be used to estimate the performance of video encoders beyond the 4^{th} generation.

Different from existing prediction models that aim to predict objective outcomes such as SSIM and PSNR scores, the proposed models have a much closer association with the end user perceived video quality on the popular small-form factor displays found on various mobile devices. In the case where objective scores are available, they can be used as additional predictors to further improve the accuracy of the prediction model. This modular design will also allow great flexibility in real life situation.

6.3 LIMITATIONS AND FUTURE RESEARCH

Due to the limited time and resources made available to this study, validation of the proposed models is not carried out. In the real world scenario, the prediction accuracy might be different from the emulated one especially when the viewing conditions and selection of participants are not controlled. If condition permits, validation of the proposed models should be carried out in a future study.

There is a possible correlation between the superior performances of the latest video encoding encoders shown and the increase of resolution resulted. This phenomenon is speculated resulted from the 64×64 large MB, which is supported by the latest generation of video encoders. However, since video contents with spatial resolution beyond full HD were not tested in this study, the correlation cannot be confirmed. It is possible that the VP9 and H.265/AVC encoders are capable of achieving even better bitrate savings as well as performance under resolutions beyond full HD, such as 4K and 8K resolutions. On the

contrary, the largest 16×16 MB defined in the previous generation of video encoders, such as H.264/AVC, will contain overly complicated encoding overhead information that takes up extra bandwidth but does not improve the quality of the video image itself. It is speculated this is the very reason that the H.264/AVC encoder performed significantly better under 720p resolution than under 1080p resolution. In this study, we have observed the upper limited in performance caused by the correlation between the largest MB supported by video encoders and the spatial resolution of video contents, and the influence of the correlation over encoder performance. Determining the best resolution H.265/AVC can work under is crucial in the sense future that mobile devices are likely to come with screen better than full HD and more 4K contents are under way in the near future. Researchers should find the optimal resolutions of the 4th generation video encoders.

Additionally, the test model HM14 encoder was the only official H.265/AVC available at the time this study was conducted to the best knowledge of the author. Later versions such as HM15 and HM16 were released when this study was accomplished. The performance implication of the amendments made in the newer versions is unknown. It was also observed during the study that the HM14 H.265/AVC test model encoder was extremely slow in encoding a given video sequence. The amount of time taken to encode a video sequence is still far from any practical use. The actual implementation of the H.265/AVC encoder will be likely to take encoding time into consideration and make compromise in the encoding quality. Therefore, the performance of the HM14 test model encoder, as discovered in this study, might not be an accurate indication of the actual performance of the implemented version of H.265/AVC. The same notion applies to the VP9 encoder as well. Although the VP9 encoder is more completed compared to the H.265/AVC encoder, the encoding time it spent on encoding the 720p and 1080p resolution video sequences is far longer than the H.264/AVC encoder will spend. Additionally, it is true that H.264/AVC encoder has already been optimized over and over again in the past, both software-wise and hardware-wise; however, the actual implementation of VP9 might change in a minor extent, to shorten the encoding time.

6.4 FURTHER DISCUSSION OF VIDEO ENCODER PERFORMANCE MODELLING

The prediction accuracy of prediction models for user perceived video quality on small-form factor screens can be enhanced in the future by better approaches to defining video content and encoders.

Full HD video contents are used in this study; however, the 4th generation video encoders might not perform the best under such resolution. The accuracy of the proposed prediction models can be improved if the best suitable resolutions for 4th generation video encoders can be discovered. Current trends in mobile hardware development and encoder design reveal 4K and 8K video contents are already going to be adopted in the near future, after full HD. Besides, one of the major reasons that the 4th generation video encoders were developed is to cater for the future need of video resolutions beyond full HD. Therefore, the correlation between video resolutions beyond full HD and the largest supported MB sizes supported by the 4th generation video encoders on mobile devices should be investigated and used as the predictor to enhance the accuracy of the prediction models. By doing so, the true potential of the 4th generation video encoders can be revealed as we have revealed the performance threshold of the H.264/AVC encoder in relation with spatial resolution in this study. Furthermore, such correlation will be able to reveal the possibility to further enhance user perceived video quality and will provide the ground truth or supporting theoretical evidence if larger MB sizes were to be introduced into the development of newer video encoders. In other word, finding this correlation is also important in the sense of developing video encoders beyond VP9 and H.265/AVC.

In the future search, the actual implementation of the H.265/AVC encoder should be used instead of the test model encoders that simulate the perfect performance as their performance might not be an accurate indication of their capability when widely used. As seen in the implementation processes of previous generation of video encoders such as H.264/AVC, in actual implementation, not all functions specified in the official specification documents of video encoders will be enabled since a balance between encoding quality and time has to be achieved for real-life situations. Due to hardware capability, compromises in video quality have often been made to improve the encoding time.

More accurate approaches to defining video contents are also needed to developed and adopted as prediction model predictors in the future. As revealed by the model tests illustrated in section 5.2, average TI and SI values do not have significant influence in the perceived video quality of end user. TI values are only used when objective scores are not available. However, both the sVQA and oVQA scores have shown that video content type influences the respective outcomes by a great extent depending on the motion intensity and image complexity of the video contents. Therefore, more accurate and reliable ways to define video contents by their level of motion and image detail should be considered in future studies.

In real-life situations, end users will use mobile devices of various screen sizes. In the future research, the influence level of mobile device screen size should be investigated as the predictor in the prediction model. This will help video content providers to customize the bitrate of the video contents and deliver them to the satisfaction of the end users.

It is expected that the prediction accuracy of the proposed models can be improved by a wide margin if video contents are more accurately defined and screen sizes of mobile devices can be considered as predictors. There is great space for the improvement of prediction accuracy of the prediction models proposed in this study.

References

Anegekuh, L., Sun, L., & Ifeachor, E. (2013). Encoding And Video Content Based HEVC Video Quality Prediction. *Multimedia Tools and Applications*, 1-24.

Apple Inc. (2014a). iPad Air Hardware Specification. from https://www.apple.com/au/ipad-air/specs/

Apple Inc. (2014b). iPhone 6 Retina HD Display. from https://www.apple.com/iphone-6/display/

Bankoski, J., Bultje, R. S., Grange, A., Gu, Q., Han, J., Koleszar, J., . . . Xu, Y. (2013). *Towards A Next Generation Open-Source Video Codec*. Paper presented at the IS&T/SPIE Electronic Imaging.

Bellard, F., & Niedermayer, M. (2012). FFmpeg.

Bjontegaard, G. (2008). Improvements of The BD-PSNR Model. *ITU-T SG16 Q, 6*, 35.

Boucher, G. (2008). VHS Era Is Winding Down. *Los Angeles Times*, A1.

Cermak, G., Pinson, M., & Wolf, S. (2011). The Relationship Among Video Quality, Screen Resolution, and Bitrate. *Broadcasting, IEEE Transactions on, 57*(2), 258-262.

Chan, M., Yu, Y., & Constantinides, A. (1990). Variable Size Block Matching Motion Compensation With Applications To Video Coding. *IEE Proceedings I (Communications, Speech and Vision), 137*(4), 205-212.

Chen, J.-W., Kao, C.-Y., & Lin, Y.-L. (2006). *Introduction To H. 264 Advanced Video Coding*. Paper presented at the Proceedings of the 2006 Asia and South Pacific Design Automation Conference.

Chikkerur, S., Sundaram, V., Reisslein, M., & Karam, L. J. (2011). Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison. *Broadcasting, IEEE Transactions on, 57*(2), 165-182.

Choi, K., & Jang, E. S. (2012). Fast Coding Unit Decision Method Based On Coding Tree Pruning For High Efficiency Video Coding. *Optical Engineering, 51*(3), 030502-030501-030502-030503.

Cisco. (2013). Global Mobile Data Traffic Forecast Update, 2012-2017. *Cisco white paper*.

CodecVisa. (2013). CodecVisa Bitstream Analyzer. from www.codecian.com

Dumic, E., Mustra, M., Grgic, S., & Gvozden, G. (2009). *Image Quality Of 4：2：2 And 4：2：0 Chroma Subsampling Formats*. Paper presented at the ELMAR, 2009. ELMAR'09. International Symposium.

Eskicioglu, A. M., & Fisher, P. S. (1995). Image Quality Measures And Their Performance. *Communications, IEEE Transactions on, 43*(12), 2959-2965.

Fan, Y.-C., Lin, H.-S., Chiang, A., Tsao, H.-W., & Kuo, C.-C. (2008). Motion Compensated Deinterlacing With Efficient Artifact Detection For Digital Television Displays. *Display Technology, Journal of, 4*(2), 218-228.

Farajzadeh, N., & Mazloumi, M. A Machine Learning Approach to No-Reference Objective Video Quality Assessment for High Definition Resources.

Feller, C., Wuenschmann, J., Roll, T., & Rothermel, A. (2011). *The VP8 Video Codec-Overview and Comparison To H. 264/AVC*. Paper presented at the Consumer Electronics-Berlin (ICCE-Berlin), 2011 IEEE International Conference on.

Fliegel, K. (2014). QUALINET Multimedia Databases v5. 0.

Frojdh, P., Horn, U., Kampmann, M., Nohlgren, A., & Westerlund, M. (2006). Adaptive Streaming Within The 3gpp Packet-Switched Streaming Service. *Network, IEEE, 20*(2), 34-40.

Garcia, R., & Kalva, H. (2013). *Human Mobile-Device Interaction On HEVC And H. 264 Subjective Evaluation For Video Use In Mobile Environment.* Paper presented at the Consumer Electronics (ICCE), 2013 IEEE International Conference on.

Gaubatz, M. Metrix MUX Visual Quality Assessment Package: MSE, PSNR, SSIM, MSSIM, VSNR, VIF, VIFP, UQI, IFC, NQM, WSNR, SNR. *htt p://foulard. ece. cornell. edu/gaubatz/metrix_mux*.

Metrix MUX Visual Quality Assessment Package: MSE, PSNR, SSIM, MSSIM, VSNR, VIF, VIFP, UQI, IFC, NQM, WSNR, SNR.

Girod, B. (1993). *What's Wrong with Mean-Squared Error?* Paper presented at the Digital images and human vision.

Gong, N., Park, C., Lee, J., Jeong, I., Han, H., Hwang, J., . . . Ha, Y. (2012). *Implementation Of 240 Hz 55—Inch Ultra Definition LCD Driven By A-Igzo Semiconductor TFT With Copper Signal Lines.* Paper presented at the Soc. Inf. Display 2012 Int. Symp. Dig. Tech. Papers.

Grois, D., Marpe, D., Mulayoff, A., Itzhaky, B., & Hadar, O. Performance Comparison Of H. 265/MPEG-HEVC, Vp9, And H. 264/MPEG-AVC Encoders.

Guo, L., & Meng, Y. (2006). *What is Wrong and Right with MSE.* Paper presented at the Eighth IASTED International Conference on Signal and Image Processing.

Gustafsson, J., Heikkila, G., & Pettersson, M. (2008). *Measuring Multimedia Quality In Mobile Networks With An Objective Parametric Model.* Paper presented at the Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on.

Hands, D., Barriac, O. V., & Telecom, F. (2008). Standardization Activities In The ITU for A QoE Assessment Of IPTV. *IEEE Communications Magazine*, 79.

Hands, D., & Brunnstrom, K. (2007). Multimedia Group Test Plan Draft Version 1.19: Video Quality Experts Group (VQEG).

Hanhart, P. (2013). VQMT: Video Quality Measurement Tool. 2014, from http://mmspg.epfl.ch/vqmt

Hanhart, P., Rerabek, M., De Simone, F., & Ebrahimi, T. (2012). *Subjective Quality Evaluation of The Upcoming HEVC Video Compression Standard.* Paper presented at the SPIE Optical Engineering+ Applications.

Hanhart, P., Rerabek, M., Korshunov, P., & Ebrahimi, T. (2013). [JCT-VC Contribution] Ahg4: Subjective Evaluation of HEVC Intra Coding For Still Image Compression: ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11.

Haskell, B. G. (1997). *Digital Video: An Introduction to MPEG-2: An Introduction to MPEG-2*: Springer.

Henot, J.-P., Ropert, M., Le Tanou, J., Kypreos, J., & Guionnet, T. (2013). *High Efficiency Video Coding (HEVC): replacing or complementing existing compression standards?* Paper presented at the Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on.

Hore, A., & Ziou, D. (2010). *Image Quality Metrics: PSNR Vs. SSIM.* Paper presented at the Pattern Recognition (ICPR), 2010 20th International Conference on.

Horowitz, M., Kossentini, F., Mahdi, N., Xu, S., Guermazi, H., Tmar, H., . . . Xu, J. (2012). *Informal Subjective Quality Comparison Of Video Compression Performance Of The HEVC And H. 264/MPEG-4 AVC Standards For Low-Delay Applications.* Paper presented at the SPIE Optical Engineering+ Applications.

Hoßfeld, T., Biedermann, S., Schatz, R., Platzer, A., Egger, S., & Fiedler, M. (2011). *The Memory Effect And Its Implications On Web QoE Modeling.* Paper presented at the Proceedings of the 23rd International Teletraffic Congress.

Huynh-Thu, Q., & Ghanbari, M. (2008). Scope of Validity Of PSNR In Image/Video Quality Assessment. *Electronics letters, 44*(13), 800-801.

Ibrahim Ali, W. (2007). *Real Time Video Sharpness Enhancement By Wavelet-Based Luminance Transient Improvement.* Paper presented at the Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on.

ITU-R, R. (1998). BT.1129-2 Subjective Assessment of Standard Definition Digital Television (SDTV) Systems.

ITU-T. (2008). P. 910 Subjective Video Quality Assessment Methods For Multimedia Applications.

Jumisko-Pyykkö, S., & Häkkinen, J. (2005). *Evaluation Of Subjective Video Quality Of Mobile Devices.* Paper presented at the Proceedings of the 13th annual ACM international conference on Multimedia.

Kamaci, N., & Altunbasak, Y. (2003). *Performance Comparison of The Emerging H. 264 Video Coding Standard With The Existing Standards.* Paper presented at the Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on.

Khan, A., Sun, L., & Ifeachor, E. (2012). QoE Prediction Model And Its Application In Video Quality Adaptation Over UMT Networks. *Multimedia, IEEE Transactions on, 14*(2), 431-442.

Khan, A., Sun, L., Ifeachor, E., Fajardo, J.-O., Liberal, F., & Koumaras, H. (2010). Video Quality Prediction Models Based on Video Content Dynamics for H. 264 Video Over UMTS Networks. *International Journal of Digital Multimedia Broadcasting, 2010*.

Khan, A., Sun, L., Ifeachor, E., Fajardo, J. O., & Liberal, F. (2010). *Video Quality Prediction Model For H. 264 Video Over UMTS Networks and Their Application In Mobile Video Streaming.* Paper presented at the Communications (ICC), 2010 IEEE International Conference on.

Kim, S. S., You, B. H., Cho, J. H., Kim, D. G., Berkeley, B. H., & Kim, N. D. (2009). An 82 - In. Ultra - Definition 120 - Hz Lcd TV Using New Driving Scheme And Advanced Super Pva Technology. *Journal of the Society for Information Display, 17*(2), 71-78.

Knoche, H., & McCarthy, J. (2004). Mobile Users" Needs And Expectations Of Future Multimedia Services.

Knoche, H., McCarthy, J. D., & Sasse, M. A. (2005). *Can Small Be Beautiful? Assessing Image Resolution Requirements For Mobile TV.* Paper presented at the Proceedings of the 13th annual ACM international conference on Multimedia.

Korhonen, J., & You, J. (2010). *Improving Objective Video Quality Assessment With Content Analysis.* Paper presented at the Proceedings of the fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM) Scottsdale, USA.

Kreis, R. (2004). Issues Of Spectral Quality In Clinical 1h - Magnetic Resonance Spectroscopy And A Gallery Of Artifacts. *NMR in Biomedicine, 17*(6), 361-381.

Lab, M. G. M. (2013). MSU Video Quality Measurement Tool.

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen Ways To Look At The Correlation Coefficient. *The American Statistician, 42*(1), 59-66.

Lim, J. S. (1998). Digital Television: Here At Last. *Scientific American, 278*(5), 78-83.

Lu, L., Wang, Z., Bovik, A. C., & Kouloheris, J. (2002). *Full-Reference Video Quality Assessment Considering Structural Distortion And No-Reference Quality Evaluation Of MPEG Video.* Paper presented at the Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on.

Martens, J.-B., & Meesters, L. (1998). Image Dissimilarity. *Signal processing, 70*(3), 155-176.

Matthews, W. J., & Stewart, N. (2009). The Effect Of Interstimulus Interval On Sequential Effects In Absolute Identification. *The Quarterly Journal of Experimental Psychology*, *62*(10), 2014-2029.

Menkovski, V., Exarchakos, G., & Liotta, A. (2010). *Machine Learning Approach for Quality Of Experience Aware Networks.* Paper presented at the Intelligent Networking and Collaborative Systems (INCOS), 2010 2nd International Conference on.

Menkovski, V., Oredope, A., Liotta, A., & Sánchez, A. C. (2009). *Predicting Quality of Experience In Multimedia Streaming.* Paper presented at the Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia.

Merritt, L., & Vanam, R. (2006). X264: A High Performance H. 264/AVC Encoder. from http://neuron2.net/library/avc/overview_x264_v8_5.pdf

Model, J. (2008). H. 264/Avc Reference Software.

Moorthy, A. K., Choi, L. K., Bovik, A. C., & de Veciana, G. (2012). Mobile Video Quality Assessment Database.

Mu, M., & Mauthe, A. (2008). *Video Quality Assessment and Management in Content Distribution Networks.* Paper presented at the Med-Hoc-Net.

Mukherjee, D., Han, J., Bankoski, J., Bultje, R., Grange, A., Koleszar, J., . . . Xu, Y. (2013). *A Technical Overview of VP9—The Latest Open-Source Video Codec.* Paper presented at the SMPTE Conferences.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied Linear Statistical Models* (Vol. 4): Irwin Chicago.

Nguyen, T., & Marpe, D. (2012). *Performance Analysis Of HEVC-Based Intra Coding For Still Image Compression.* Paper presented at the Picture Coding Symposium (PCS), 2012.

Nightingale, J., Wang, Q., Grecos, C., & Goma, S. (2013). Modeling QoE For Streamed H. 265/HEVC Content Under Adverse Network Conditions.

NTT. Video Quality Assessment Methods.

Ohm, J.-R., Sullivan, G. J., Schwarz, H., Tan, T. K., & Wiegand, T. (2012). Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC). *IEEE Trans. Circuits and Systems for Video Technology, 22*(12), 1669-1684.

Ohm, J., Sullivan, G. J., Schwarz, H., Tan, T. K., & Wiegand, T. (2012). Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC). *Circuits and Systems for Video Technology, IEEE Transactions on, 22*(12), 1669-1684.

Ooyala. (2013). Q1 2013 Video Index - TV Is No Longer A Single Screen In Your Living Room.

Péchard, S., Pépion, R., & Le Callet, P. (2008). *Suitable Methodology In Subjective Video Quality Assessment: A Resolution Dependent Paradigm.* Paper presented at the Proceedings of the Third International Workshop on Image Media Quality and its Applications, IMQA2008.

Pinson, M. H., & Wolf, S. (2004). A New Standardized Method For Objectively Measuring Video Quality. *Broadcasting, IEEE Transactions on, 50*(3), 312-322.

Pinson, M. H., Wolf, S., & Cermak, G. (2010). HDTV Subjective Quality Of H. 264 Vs. MPEG-2, With And Without Packet Loss. *Broadcasting, IEEE Transactions on, 56*(1), 86-91.

Pourazad, M. T., Doutre, C., Azimi, M., & Nasiopoulos, P. (2012). HEVC: The New Gold Standard For Video Compression: How Does HEVC Compare With H. 264/AVC? *Consumer Electronics Magazine, IEEE, 1*(3), 36-46.

Protalinski, E. (2013). Google Adds Its Free And Open-Source VP9 Video Codec To Latest Chrome Build.

Raake, A., Garcia, M.-N., Moller, S., Berger, J., Kling, F., List, P., . . . Heidemann, C. (2008). *TV-Model: Parameter-Based Prediction of IPTV Quality.* Paper presented at the Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on.

Ramsey, P. H. (1989). Critical Values for Spearman's Rank Order Correlation. *Journal of Educational and Behavioral Statistics, 14*(3), 245-253.

Rao, K. R. (2013). *Video Coding Standards: AVS China, H. 264/MPEG-4 part 10, HEVC, VP9, DIRAC and VC-1.* Paper presented at the Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2013.

Rec, I. (2008). P. 910 Subjective Video Quality Assessment Methods For Multimedia Applications.

Rec, I. (2012a). BT. 500-13 Methodology For The Subjective Assessment Of The Quality Of Television Pictures.

Rec, I. (2012b). BT. 1210-4 Test Materials To Be Used In Assessment Of Picture Quality.

Rerabek, M., & Ebrahimi, T. (2014). *Comparison Of Compression Efficiency Between HEVC/H. 265 And Vp9 Based on Subjective Assessments.* Paper presented at the SPIE Optical Engineering+ Applications.

Řeřábek, M., & Ebrahimi, T. (2014). *Comparison Of Compression Efficiency Between HEVC/H. 265 And Vp9 Based On Subjective Assessments.* Paper presented at the SPIE Optical Engineering+ Applications.

Richardson, I. E. (2004). *H. 264 And MPEG-4 Video Compression: Video Coding for Next-Generation Multimedia*: John Wiley & Sons.

Ries, M., Nemethova, O., & Rupp, M. (2007). *Performance Evaluation Of Mobile Video Quality Estimators.* Paper presented at the Proceedings of the European Signal Processing Conference,(Poznan, Poland.

Rohaly, A. M., Corriveau, P. J., Libert, J. M., Webster, A. A., Baroncini, V., Beerends, J., . . . Harrison, D. (2000). *Video Quality Experts Group: Current Results and Future Directions.* Paper presented at the Visual Communications and Image Processing 2000.

Sasse, M. A., & Knoche, H. (2006). *Quality in Context-An Ecological Approach To Assessing QoS For Mobile TV.* Paper presented at the Proceedings of 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems.

Savakis, A. E., Etz, S. P., & Loui, A. C. (2000). *Evaluation Of Image Appeal In Consumer Photography.* Paper presented at the Electronic Imaging.

Sayood, K. (2002). Statistical Evaluation of Image Quality Measures. *Journal of Electronic imaging, 11*(2), 206-223.

Schylander, E. (1998). *Digital Video On Compact Disk.* Paper presented at the Photonics China'98.

Seshadrinathan, K., & Bovik, A. C. (2009). *Motion-Based Perceptual Quality Assessment of Video.* Paper presented at the IS&T/SPIE Electronic Imaging.

Seshadrinathan, K., & Bovik, A. C. (2010). Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos. *Image Processing, IEEE transactions on, 19*(2), 335-350.

Seshadrinathan, K., Soundararajan, R., Bovik, A. C., & Cormack, L. K. (2010). Study Of Subjective and Objective Quality Assessment Of Video. *Image Processing, IEEE transactions on, 19*(6), 1427-1441.

SGP-NewsMan, S.-P. c. (2013). Samsung Galaxy S4 Singapore Prices, Specs & Features 18 Apr 2013 Singapore.

Sikora, T. (1997). MPEG Digital Video-Coding Standards. *Signal Processing Magazine, IEEE, 14*(5), 82-100.

Smith, J. R. (2006). The H. 264 Video Coding Standard. *IEEE Computer Society, 13*, 86-90.

Song, W. (2012). User-Driven Quality Of Experience Modelling for Mobile Video Optimisation.

Song, W., Tjondronegoro, D., & Docherty, M. (2010). Exploration And Optimization Of User Experience In Viewing Videos On A Mobile Phone. *International Journal of Software Engineering and Knowledge Engineering, 20*(08), 1045-1075.

Song, W., Tjondronegoro, D. W., & Docherty, M. (2012). Understanding User Experience of Mobile Video: Framework, Measurement, and Optimization. *Mobile Multimedia: User and Technology Perspectives*, 3-30.

Standardisation, I. O. F. (2013). ISO/IEC JTC 1/SC 29/WG 11 - Coding Of Moving Pictures And Audio.

Sullivan, G. J., Ohm, J., Han, W.-J., & Wiegand, T. (2012). Overview Of The High Efficiency Video Coding (HEVC) Standard. *Circuits and Systems for Video Technology, IEEE Transactions on, 22*(12), 1649-1668.

Sullivan, G. J., Topiwala, P. N., & Luthra, A. (2004). *The H. 264/Avc Advanced Video Coding Standard: Overview And Introduction To The Fidelity Range Extensions.* Paper presented at the Optical Science and Technology, the SPIE 49th Annual Meeting.

Teo, P. C., & Heeger, D. J. (1994). *Perceptual Image Distortion.* Paper presented at the IS&T/SPIE 1994 International Symposium on Electronic Imaging: Science and Technology.

Vatolin, D., Kulikov, D., Parshin, A., Titarenko, A., & Soldatov, S. (2007). MPEG-4 AVC/H. 264 Video Codecs Comparison. *CS MSU Graphics & Media Lab*.

Vranjes, M., Rimac-Drlje, S., & Zagar, D. (2008). *Subjective And Objective Quality Evaluation of The H. 264/AVC Coded Video*. Paper presented at the Systems, Signals and Image Processing, 2008. IWSSIP 2008. 15th International Conference on.

Wang, Z., Bovik, A. C., & Evan, B. (2000). *Blind Measurement Of Blocking Artifacts In Images.* Paper presented at the Image Processing, 2000. Proceedings. 2000 International Conference on.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility To Structural Similarity. *Image Processing, IEEE transactions on, 13*(4), 600-612.

Wang, Z., Sheikh, H. R., & Bovik, A. C. (2003). Objective Video Quality Assessment. *The handbook of video databases: design and applications*, 1041-1078.

Wang, Z., & Simoncelli, E. P. (2005). *Reduced-Reference Image Quality Assessment Using A Wavelet-Domain Natural Image Statistic Model.* Paper presented at the Electronic Imaging 2005.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). *Multiscale Structural Similarity for Image Quality Assessment.* Paper presented at the Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on.

Webster, A. A., Jones, C. T., Pinson, M. H., Voran, S. D., & Wolf, S. (1993). *Objective Video Quality Assessment System Based On Human Perception.* Paper presented at the IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology.

Wiegand, T., Sullivan, G. J., Bjontegaard, G., & Luthra, A. (2003). Overview Of The H. 264/AVC Video Coding Standard. *Circuits and Systems for Video Technology, IEEE Transactions on, 13*(7), 560-576.

Wilkinson, A. (2014). Internet Speeds: As the Gap Widens. Retrieved from http://www.webanalyticsworld.net/2014/08/internet-speed-gap-widens.html

Winkler, S. (1999). *Perceptual Distortion Metric For Digital Color Video.* Paper presented at the Electronic Imaging'99.

Winkler, S. (2012). Analysis Of Public Image And Video Databases For Quality Assessment. *Selected Topics in Signal Processing, IEEE Journal of, 6*(6), 616-625.

Wolf, S., & Pinson, M. (2007). *Application Of The Ntia General Video Quality Metric (VQM) To HDTV Quality Monitoring.* Paper presented at the Proceedings of The Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, AZ, USA.

Wong, A. H., & Chen, C.-T. (1993). *Comparison Of Iso MPEG1 And MPEG2 Video-Coding Standards.* Paper presented at the Visual Communications' 93.

Yang, W. (2008). Poly NYU Video Quality Database.

Yendrikhovski, S. N., Blommaert, F. J., & de Ridder, H. (1998). *Perceptually Optimal Color Reproduction.* Paper presented at the Photonics West'98 Electronic Imaging.

Yuen, M., & Wu, H. (1998). A Survey Of Hybrid MC/DPCM/DCT Video Coding Distortions. *Signal processing, 70*(3), 247-278.

Zhu, K., Asari, V., & Saupe, D. (2013). *No-Reference Quality Assessment of H. 264/AVC Encoded Video Based On Natural Scene Features.* Paper presented at the SPIE Defense, Security, and Sensing.